Vol.2013-MPS-96 No.22
Vol.2013-BIO-36 No.22
2013/12/12

モンテカルロ法を用いた進化分岐図作成法

石井一夫[1]*、松田朋子[2]、古崎利紀[1]、後藤哲雄[2]（東京農工大学[1]、茨城大学[2]）

*東京農工大学 農学系ゲノム科学人材育成プログラム
〒183-8509 東京都府中市幸町 3-5-8 Email: kishii@cc.tuat.ac.jp

要旨

進化系統樹（分岐図）は、一個ないし数個の遺伝子やタンパク質の配列の類似性をもとに作成されてきた。次世代シーケンサーの普及により、ゲノムスケールで新規のアセンブリにより得られた多数の遺伝子やタンパク質の組合せにより、分岐図を作成することが可能になった。しかし、その組合せパターンは無数にあり、その組合せにより微妙に異なる分岐図ができるが、すべての組合せで分岐図を検討するのは実際には不可能である。そこで、モンテカルロ法（ブートストラップ法）により、無作為な組合せの遺伝子を選択し、並列分散処理により多数回の分岐図作成を繰り返すことによる進化的分岐図作成の最適化の検討を行なった。この方法によりどの遺伝子が分岐図作成に大きく寄与しているかを評価することも可能となった。

１．はじめに

次世代シーケンサーから産生された配列データをアセンブリ（連結）して得たコンティグを元に、ゲノム配列や、遺伝子配列の網羅的データを用いて分岐図（進化系統樹）を作成すること試みた。分岐図は、核ゲノム上の 18S rRNA やミトコンドリアゲノムなどの配列を元に作成されてきたが、次世代シーケンサーからのデータにより全ゲノム配列レベルでの分岐図を作成し、包括的なデータを用いた進化の解析が可能になる。しかし、全ゲノム配列を用いて分枝図を作成するのは、ゲノム配列のバリアントの多さ、偏在性、それらの組合せの数などを考えると非常に膨大な組合せの配列データが存在する。今回、RNA-Seq（次世代シーケンサーを用いた網羅的発現解析）により得られた遺伝子配列データを用いて、それらの組合せ全体から無作為に遺伝子の組合せを抽出して、RNA-Seq（次世代シーケンサーを用いた網羅的発現解析）により得られたに遺伝子並列データを用いて、マルチプルアラインメントを行ない、分岐図を作成するという操作を繰り返し、その頻度で最適化された分岐図を作成することを試みた。

Vol.2013-MPS-96 No.22
Vol.2013-BIO-36 No.22
2013/12/12

2. 材料および方法

分岐図作成に用いた遺伝子群：

次世代シーケンサーHi-Seq を用いて 28 種類の近縁生物種の網羅的発現解析(RNA-Seq)を行いアセンブリソフトVelvet及びOases を用いてできた約数万のコンティグの相互の類似性をもとに選択した 447 種の遺伝子配列を用いて MAFFT と trimAL を用いて作成したマルチプルアラインメントのファイルを用いて検討した。

まず、28 種類の近縁生物種の生物（ゲノムサイズ約 100M 塩基対）から総 RNA を精製し、メッセンジャーRNA の網羅的配列解析を行なった。各検体について、約 100～150 塩基長、約 2000 万～5000 万リードの配列データを得た。これを、個々にアセンブリ（整列編集）作業を行い、それぞれ数万のコンティグ（塩基配列の塊）を得た。これを、マルチプルアラインメントソフトウェア MAFFT と TrimAL を用いて多重整列（マルチプルアラインメント）を行なった。

組合せ数列の作成とモンテカルロ法による選択：

447 個の遺伝子から 1 遺伝子を選択する組合せの数は、R-3.0.2 により以下のように計算し、447 個存在する。

```
> choose(447,1)  # 447 個の遺伝子から 1 遺伝子を選択
[1] 447
```

447 個の遺伝子から 2 遺伝子までを選択する組合せの数は、以下のように計算し、100,128 個存在する。また、447 個の遺伝子から 3 遺伝子まで、4 遺伝子までを選択する組合せの数は、以下のように計算し、それぞれ 14,886,143 個、1,656,133,808 個存在する。

```
> choose(447,1)+choose(447,2) # 2 遺伝子を選択
[1] 100128
> choose(447,1)+choose(447,2)+choose(447,3)
[1] 14886143                                  # 3 遺伝子を選択
> choose(447,1)+choose(447,2)+choose(447,3)+
choose(447,4)                                 # 4 遺伝子を選択
[1] 1656133808 <- 今回はこれで検証(該当する遺伝子を除く)
```

Vol.2013-MPS-96 No.22
Vol.2013-BIO-36 No.22
2013/12/12

447 個の遺伝子から 1 個から 447 個まですべてを選択する組合せの数は、以下のように計算し、6.109568e+99 個存在する。

```
> choose(447,1)+choose(447,2)+choose(447,3)+.....+choose(447,447) # 447遺伝子を選択
[1] 6.109568e+99
```

ここでは、6.109568e+99 とおりの全ての場合の数に対応する遺伝子に対して分岐図を作成することは現実として、不可能であるので、447 個の遺伝子から、4 遺伝子までを選択する場合の数に相当する 1,656,133,808 個の遺伝子の組合せの中から、モンテカルロ法によりランダムに対応する遺伝子を除き、分岐図作成ソフト RAxML を用いて分岐図作成を繰り返し、最適な分岐図を作成することとした。

447 個の組合せから 4 個の数列を選択する 1656133808 個の組合せ数列は、以下のような Perl スクリプトを用いて作成した。

```
#!/usr/local/bin/perl
use strict;
use warnings;
our $n = 447;
our $f = 0;
flex('', 1, $n) if $f;
for(my $r = 1; $r <= $n; $r++){
    flex('', 1, $r);
    }
sub flex{
    my($c, $begin, $r) = @_;
    if($r == 1){
            for(my $i = $begin; $i <= $n; $i++){
                    print "$c$i¥n";
            }
          return;
        }else{
                $r--;
```

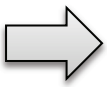
Vol.2013-MPS-96 No.22
Vol.2013-BIO-36 No.22
2013/12/12

```
                my $end = $n - $r;
                for(my $i = $begin; $i <= $end; $i++){
                        flex("$c$i,", $i + 1, $r);
                }
        }
}
```

次に、R により以下のように 1656133808 個の整数から一様乱数を発生させ、それに対応する番号を上記の Perl スクリプトで作成した組合せ数列から選択した。

```
> run1 <- round(runif(100000, min = 1, max = 1656133808))
```

447 個の遺伝子からこれに対応する番号の遺伝子を除いたマルチプルアラインメントを用いて、RAxML により分岐図を作成した。

発生させた乱数		乱数に対応する数列
81571247		5, 121, 150, 372
365913044		27, 29, 176, 229
1023396239		95, 160, 302, 413
644416555		51, 199, 232, 430
….		….

選択したアラインメントによる分岐図作成とテキストパターンへの変換：

マルチプルアラインメントは、EMBOSS-6.4.0 の seqret 関数を用いて fasta 形式ファイルから phylip 形式ファイルへ変換し、RAxML により分岐図を作成した。得られた分岐図は、以下に示すような分枝の長さを示す数値や記号が入っており以下のような複雑な出力結果であるが、これらを除き単純なテキストパターンに変換した。

==> RAxML_bestTree.out <==　# RAxML の出力結果例

(((x:0.00531404813722460411,b:0.00582680929801783314):0.025971266250438409
39,(m:0.03645324955994154459,((i:0.02130122160079299734,g:0.031447907657455
42060):0.00424151281867054565,u:0.02948686769656536782):0.020403600469400
21752):0.015217749948996110003):0.00840806219060770237,((((z:0.3434476718995
4156228,(t:0.14262613954560879326,l:0.096110243065704065517):0.124827271887

54781655):0.17738679389459199864,(((q:0.09220903636333667441,c:0.0784974040
2964605623):0.02162213710044687265,(((((e:0.02209622018870339294,k:0.028932
83119099681472):0.008971545350474785 52,p:0.08526167426794276083):0.013931
93949473354662,(f:0.01318356630444799359,(n:0.01952490523202302791,&:0.016
93127308779425119):0.008136049288154000 03):0.07817988855466992404):0.0226
3432315084732208,(r:0.04590445767519640841,h:0.04637315388999797144):0.038
65318679664145329):0.00789191373814705256,(v:0.03284611917996409919,d:0.02
25021056831853257 0):0.17923032798411692168):0.01001049487098192720):0.103
25734189742048763,(y:0.20086249354886381857,(j:0.18321752975378832740,#:0.2
09679853903260327 02):0.12992858329809614526):0.01880523845322217696):0.03
85741459174890263 8):0.06620101031901222399,(o:0.03854648520093560682,s:0.0
35527049274526989 46):0.12706855170728659221):0.05630830036902149949,w:0.1
39336296263945474 96):0.07157780409461377003,a:0.037820217201590664 02):0.0;

==> RAxML_bestTreeSUMMARY.out <== # 変換された簡素化分岐図パターン
(((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),(((((e,k),p),(f,(n,&))),(r,h)),(v,d))),(y,(j,#)))),(o,s)),w)
,a)

3、試行結果

無作為抽出したマルチプルアラインメントによる分岐図作成を何回も繰り返し、集計することにより最適化分岐図を得た。合計 200 回の試行を行ない 185 回が以下のようなメジャーな分岐図パターンで、15 回がマイナーな分岐図パターンであった。

メジャーな分岐図パターン：
(((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),(((((e,k),p),(f,(n,&))),(r,h)),(v,d))),(y,(j,#)))),(o,s)),w)
,a)

マイナーな分岐図パターン：
(((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((v,d),((((e,k),p),(f,(n,&))),(r,h)))),(y,(j,#)))),(o,s)),w)
,a)

200 回の試行のうちメジャーな分岐図メジャーな分岐図パターンで除かれた遺伝子は分岐図作成への寄与が小さい遺伝子（357 遺伝子）で、マイナーな分岐図パターンで

Vol.2013-MPS-96 No.22
Vol.2013-BIO-36 No.22
2013/12/12

除かれた遺伝子は分岐図作成への寄与が高い遺伝子（57 遺伝子）であると考えられた。両方のパターンで除かれる遺伝子は、単独のパターンで除かれる遺伝子より、メジャーな分岐図パターンで除かれた遺伝子の場合、寄与度はより高く、マイナーな分岐図パターンで除かれた遺伝子の場合、寄与度はより低いと考えられた。マイナーな分岐図パターンでのみ除かれた分岐図作成に寄与度の高いと考えられる遺伝子として、遺伝子 105、106（以上、頻度 2 回）、7、40、42、77、267、283、331、344、388、411（以上、頻度 1 回）が同定された。また、メジャーな分岐図パターンで複数回除かれた分岐図作成に寄与度の低いと考えられる遺伝子として、108、167（以上、頻度 6 回）、57、71、78、174、180、326、341、401（以上、頻度 7 回）などがリストアップされた。

4、まとめと結論

モンテカルロ法によるブートストラップで無作為に抽出した遺伝子群の配列情報をもとに最適な進化系統樹（分岐図）を作成する方法を確立した。また、抽出した遺伝子の頻度の解析から、分岐図作成に寄与度の高い遺伝子、寄与度の低い遺伝子をリストアップすることが可能であった。

謝辞

本研究は、文部科学省特別経費「農学系ゲノム科学領域における人材育成プログラム」の支援により実施した。本人材育成プログラムを通してご支援いただき貴重なコメントをいただいたプログラム代表の東京農工大学教授高橋信弘先生、特任教員、その他のご支援頂いた教員の先生方、プログラムに参加した学生諸氏に感謝いたします。また、本研究は日本ヒューレットパッカード（株）の計算機リソース提供と、Amazon Web Services, Inc.の AWS in Education Grant award の支援により実施した。

参考文献

1) Rizzo M. (2008) Statistical Computing with R, Chapman & Hall/CRC (Rizzo M(著) , 石井一夫, 村田真樹（共訳）(2011)『R よる計算機統計学』オーム社)
2) 石井一夫、佐藤暁、古崎利紀、有江力、寺岡徹、ゲノム科学におけるビッグデータ・データマイニング、日本統計学会誌、43 巻 1 号 90-111 頁 (2013)