羊膜類の生物進化系統樹の推定

00-11593
嶋村謙太

00-16188
富岡さやか

2月10日

1 ABST

羊膜類の各種動物のDNAアミノ酸配列データから、最尤検定法等を用いて進化系統樹を推測した。鳥類、ハ虫類、両性類の進化系統部分の結果に疑問を感じたので、そこを推測し直したところ、現在信頼されている系統樹と違う系統樹が高い確率で正しいという結果が得られたが、現在信頼されている系統樹も棄却はされなかった。

2 INTRO

下平先生が出された課題で、各種動物について最尤法などを用いて進化系統樹を推測することになったので、名前を知っている動物の中から幅広くサンプルをとって進化系統樹を推測した。その後、系統樹の推定結果のカメの位置が、現在信頼されていると思っていた系統樹での位置と違ったので、データの洗いなおしをし、その部分のサンプルを多くとりなおして、検定をしなおした。

3 Methods

3.1 概要

NCBIのデータベースからDNA塩基配列部分を得る。ミトコンドリアゲノムのファイルを23個ダウンロードした。そうして得られたデータを(*)NJ法，最尤法，の順で行う。最尤法で得られた系統樹において不確定と思われる部分木を再考する。枝の長さをゼロとして、その部分のあらゆる部分木の組み合わせを考え、それらの組み合わせの確率をそれぞれ求める。その中で，auの値が棄却できない統計樹，枝に注目する．(*)気になった点(カメとイグアナ)をもとに再びデータを構成しなおす．データベースからミトコンドリアゲノムのファイルを新しく5個追加しホ乳類を3個にまで減らす．それらのデータ14個を用いて再び(*)～(*)を行う．

3.2 23個のデータをもとに実行

3.2.1 データのダウンロード

NCBIのデータベースからDNA塩基配列部分を得る。NCBIのデータベースからミトコンドリアゲノムのファイルをダウンロードしそのファイルの中から遺伝子のDNA配列だけを取り出す。．ダウンロードファイルは以下の生物のデータである．

NC000845:Sus scrofa:pig:ブタ
NC000846:Rhea americana:greater rhea:レア
NC000884:Cavia porcellus:domestic guinea pig:モルモット
NC000886:Chelonia mydas:green seaturtle:アオウミガメ
NC000891:Ornithorhynchus anatinus:platypus:カモノハシ
NC000934:Loxodonta africana:African savanna elephant:アフリカゾウ
NC001573:Xenopus laevis:African clawed frog:アフリカツメガエル
NC001601:Balaenoptera musculus:blue whale:シロナガスクジラ
NC001602:Halichoerus grypus:gray seal:ハイイロアザラシ
NC001610:Didelphis virginiana:North American opossum:キタオポッサム
NC001640:Equus caballus:horse:ウマ
NC001644:Pan paniscus:pygmy chimpanzee:ピグミーチンパンジー
NC001645:Gorilla gorilla:ゴリラ
NC001807:Homo sapiens:human:ヒト
NC001922:Alligator mississippiensis:American alligator:アメリカアリゲーター

NC002081:Gadus morhua:Atlantic cod:タイセイヨウダラ
NC002082:Hylobates lar:common gibbon:シロテテナガザル
NC002083:Pongo pygmaeus abelii:Sumatran orangutan:スマトラオランウータン
NC002746:Isoodon macrourus:northern brown bandicoot:シモフリコミミバンディクート
NC002793:Iguana iguana:common iguana:イグアナ
NC003321:Tachyglossus aculeatus:Australian echidna:ハリモグラ
NC003426:Ursus americanus:American black bear:アメリカクロクマ
NC004028:Lepus europaeus:European hare:ノウサギ

3.2.2 NJ 法

はじめに outgroup(最も遠縁な生物) を NC000845:ブタにして NJ 法を行い，さらにその結果から outgroup を NC002081:タイセイヨウダラとして NJ 法を行う。

3.2.3 最尤法

実行したNJ法で得られた系統樹をもとに，さらに信頼性を高めるために最尤法で新たな系統樹を得る。ここではNJ法で得た系統樹を尤度を大きくするように系統樹を少しずつ変更していき，最大の尤度になったらそれを最尤推定の近似とみなすという方法をとる。そうして得られた系統樹の枝のローカルブートストラップ確率を参考に確率の小さい不確定な部分と，比較的確定的と思われる部分とに分ける。不確定な部分の枝の長さをゼロに縮め，その部分から作られる全ての部分系統樹の組み合わせを考える。
ここで展開する時に，Fig.1に従って大きく2つのグループに分ける。A1、B1、C1、D1、E1を不確定にしたもの(αとする．A2、B2、C2、D2、E2は固定)とA2、B2、C2、D2、E2を不確定にしたもの(βとする．A1、B1、C1、D1、E1は固定)とわけて別々に実行しこれらの組み合わせを考えた。

そうして得られた全ての系統樹について、それぞれ尤度を求め、系統樹の確率値を計

算する。また同様にそれぞれの枝についても確率値を計算する。系統樹の確率値のなかでau値に着目する．αを不確定にして求めた系統樹の1位，2位，3位とβを不確定にして求めた1位と2位を組み合わせた
(α順位，β順位)=(1，1)(2，1)(3，1)(1，2)
の4通りを考える．これらの各々についてNJ法，最尤法を行い，そのなかで$\ln L$の値が最も大きいものを選び，再び系統樹の確率を求めた。
そうして得られた結果はカメ，イグアナの関係がデータベース(NCBI)の形と異なることがわかったので，それらをもっと詳しく調べるために再びデータを構成しなおす。

3.3 構成しなおしたデータをもとに実行

3.3.1 データのダウンロード

同様にNCBIのデータベースから以下の種類の生物のミトコンドリアゲノムをダウンロードする。

NC000846:Rhea americana:greater rhea:レア
NC000886:Chelonia mydas:green seaturtle:アオウミガメ
NC000888:Eumeces egregius:mole skink:トカゲの一種
NC000891:Ornithorhynchus anatinus:platypus:カモノハシ
NC001573:Xenopus laevis:African clawed frog:アフリカツメガエル
NC001640:Equus caballus:horse:ウマ
NC001807:Homo sapiens:human:ヒト
NC001922:Alligator mississippiensis:American alligator:アメリカアリゲータ

NC001945:Dinodon semicarinatus:colubrid snake:アカマタ
NC001947:Pelomedusa subrufa:helmeted turtle:ヌマヨコクビガメ
NC002073:Chrysemys picta:painted turtle:アカセスジニシキガメ
NC002081:Gadus morhua:Atlantic cod:タイセイヨウダラ
NC002780:Dogania subplana:Malayan softshell turtle:ヒラタスッポン

NC002793:Iguana iguana:common iguana:イグアナ

3.3.2 NJ 法

はじめから outgroup を NC002081:タイセイヨウダラとして NJ 法を行う。結果より，outgroup がこのままでいいことを確認できる。

3.3.3 最尤法

先程と同様に NJ 法のデータをもとに最尤法を行う。最尤推定の近似のより求めた系統樹 Fig.2 について不確定な部分を A,B,C,D,E とし、全ての系統樹 105 通りについてそれぞれ尤度を求め、系統樹の確率値を計算する。こうして得られた結果から au 値が 0.05 以下を棄却すると残る系統樹は 6 通りであった。また同様にそれぞれの枝についても確率値を計算すると棄却されない枝は 10 通りであった。

4 Results

4.1 23 個のデータをもとに実行

4.1.1 NJ 法

初めに outgroup をブタにして NJ 法を行う。ここでは他の生物と最も遠い親戚にあたる生物をタイセイヨウダラとして選んだので、タイヨウセイダラが直接ついている枝の番号を読み取りそれを outgroup として再び NJ 法を行う。得られた結果は Fig.3 である。

4.1.2 最尤法

ここで更に信頼性を高めるため新しい系統樹を得る．不確定な部分の枝の長さをゼロに縮め，その部分から作られる全ての部分系統樹の組み合わせを考える。
Fig.1 のように A1、B1、C1、D1、E1 を不確定にしたものと A2、B2、C2、D2、E2 を不確定にしたものとわけをそれぞれ二人で分担して実行した．

つまりこのような2パターンで行った.

- 富岡 : (((((A1、B1、C1、D1、E1F1)B2)C2)D2)E2)
- 嶋村 : A2、B2、C2、D2、E2

その結果求められたのが次のデータである.

- 富岡 : (((((A1、B1、C1、D1、E1F1)B2)C2)D2)E2)

```
# reading work-conc-tree.pv
#   0    1      2      9+     10  |     3      4      5      6      7      8
# rank item    obs     au     np  |    bp     pp     kh     sh    wkh    wsh |
#    1    4   -4.4  0.727  0.397  | 0.399  0.988  0.606  0.952  0.606  0.954 |
#    2   15    4.4  0.515  0.260  | 0.251  0.012  0.394  0.851  0.394  0.845 |
#    3    3   10.2  0.429  0.245  | 0.249  4e-05  0.348  0.736  0.348  0.796 |
#    4    5   15.0  0.210  0.062  | 0.062  3e-07  0.201  0.631  0.201  0.630 |
#    5   12   16.9  0.139  0.033  | 0.032  4e-08  0.121  0.590  0.121  0.454 |

# reading work-conc-edge.pv
#   0    1      2      9+     10  |     3      4      5      6      7      8
# rank item    obs     au     np  |    bp     pp     kh     sh    wkh    wsh |
#    1    6  -10.2  0.716  0.686  | 0.682  1.000  0.652  0.941  0.652  0.928 |
#    2    4   -4.4  0.643  0.459  | 0.461  0.988  0.606  0.911  0.606  0.921 |
#    3   10    4.4  0.468  0.266  | 0.256  0.012  0.394  0.812  0.394  0.798 |
#    4    5   10.2  0.397  0.248  | 0.252  4e-05  0.348  0.707  0.348  0.755 |
#    5    1   10.2  0.332  0.307  | 0.313  4e-05  0.348  0.697  0.348  0.729 |
#    6    9   16.9  0.138  0.033  | 0.032  4e-08  0.121  0.575  0.121  0.429 |
```

- 嶋村 : A2、B2、C2、D2、E2

```
# reading work-conc-tree.pv
#   0    1      2      9+     10  |     3      4      5      6      7      8
# rank item    obs     au     np  |    bp     pp     kh     sh    wkh    wsh |
#    1    5  -21.2  0.948  0.752  | 0.747  1.000  0.913  0.999  0.877  0.998 |
#    2    7   30.1  0.207  0.085  | 0.089  9e-14  0.123  0.311  0.123  0.448 |
#    3    2   21.2  0.165  0.049  | 0.047  6e-10  0.087  0.491  0.087  0.350 |
#    4   11   36.9  0.157  0.047  | 0.048  9e-17  0.094  0.207  0.094  0.377 |
#    5    3   23.8  0.099  0.034  | 0.034  5e-11  0.062  0.441  0.062  0.265 |
#    6    6   28.7  0.073  0.034  | 0.033  3e-13  0.055  0.340  0.055  0.241 |
```

```
# reading work-conc-edge.pv
#   0    1      2      9+     10  |     3      4      5      6      7      8
# rank item    obs     au     np  |    bp     pp     kh     sh    wkh    wsh |
#    1    1  -21.2  0.918  0.785  | 0.779  1.000  0.913  0.991  0.877  0.992 |
#    2    4  -28.7  0.909  0.834  | 0.828  1.000  0.945  0.993  0.877  0.987 |
#    3    2   30.1  0.205  0.085  | 0.089  9e-14  0.123  0.309  0.123  0.408 |
#    4    8   30.1  0.145  0.133  | 0.139  9e-14  0.123  0.305  0.123  0.392 |
#    5    3   21.2  0.134  0.095  | 0.096  6e-10  0.087  0.472  0.094  0.343 |
#    6    5   23.8  0.099  0.034  | 0.034  5e-11  0.062  0.433  0.062  0.249 |
#    7    7   28.7  0.072  0.034  | 0.033  3e-13  0.055  0.333  0.055  0.219 |
```

αを不確定にして求めた系統樹の1位，2位，3位とβ部を不確定にして求めた1位と2位を組み合わせた
(α順位，β順位)=(1，1)(2，1)(3，1)(1，2)
の4通りを考える．それぞれをNJ法，最尤法で実行しその中で値が最も大きいものを選ぶ．NJ法の段階で(1，1)(2，1)は一致しそれが最も値が大きいものとなった．そうして得られた系統樹で棄却できなかったものはFig.4であるそ．うして得られた結果の1位がカメ，イグアナの関係がデータベース(NCBI)の形と異なることがわかったので，それらをもっと詳しく調べるために再びデータを構成しなおす。.

4.2 構成しなおしたデータをもとに実行

4.2.1 NJ法

NJ法により得られた系統樹はFig.4である。

4.2.2 最尤法

得られたFig.2を元に，不確定な部分を でくくると次のようになる。
((A,B,C,D,E,F),G)
そうして得られた結果は木，枝とも以下のようになる．

```
# reading work-conc-tree.pv
#   0    1      2      9+     10  |     3      4      5      6      7      8
# rank item    obs     au     np  |    bp     pp     kh     sh    wkh    wsh |
#    1    1  -15.9  0.932  0.620  | 0.613  1.000  0.858  1.000  0.831  1.000 |
#    2   13   28.4  0.266  0.110  | 0.110  5e-13  0.169  0.608  0.169  0.711 |
#    3    9   15.9  0.242  0.098  | 0.100  1e-07  0.142  0.820  0.142  0.663 |
#    4   14   31.1  0.210  0.072  | 0.072  3e-14  0.145  0.563  0.145  0.662 |
#    5    6   17.7  0.175  0.069  | 0.072  2e-08  0.107  0.795  0.107  0.588 |
```

```
#    6    4   30.2  0.075  0.030 |  0.029  8e-14  0.046  0.605  0.046  0.364 |
#    7   12   40.0  0.038  0.003 |  0.003  4e-18  0.071  0.422  0.071  0.470 |

# reading work-conc-edge.pv
#   0     1      2      9+     10 |     3      4      5      6      7      8
# rank item    obs     au     np |    bp     pp     kh     sh    wkh    wsh |
#    1    2  -15.9  0.882  0.654 |  0.646  1.000  0.858  0.996  0.831  0.998 |
#    2    1  -28.4  0.873  0.785 |  0.786  1.000  0.831  0.999  0.831  0.997 |
#    3    3  -28.4  0.820  0.814 |  0.815  1.000  0.831  0.998  0.831  0.995 |
#    4    8   28.4  0.255  0.111 |  0.110  5e-13  0.169  0.593  0.169  0.661 |
#    5   16   15.9  0.241  0.098 |  0.100  1e-07  0.142  0.804  0.142  0.628 |
#    6   10   31.1  0.201  0.072 |  0.072  3e-14  0.145  0.555  0.145  0.622 |
#    7    6   28.4  0.191  0.185 |  0.185  5e-13  0.169  0.586  0.169  0.631 |
#    8   18   28.4  0.181  0.186 |  0.185  5e-13  0.169  0.596  0.169  0.619 |
#    9   19   17.7  0.174  0.069 |  0.072  2e-08  0.107  0.780  0.107  0.556 |
#   10   21   30.2  0.074  0.030 |  0.029  8e-14  0.046  0.596  0.046  0.341 |
#   11   25  136.3  0.005 6e-05 |      0  6e-60  1e-04  3e-04      0  2e-04 |
```

木は棄却できないものが6個，枝は10個のこる．確率の高いものから並べると，Fig.5が残る。 このうち最も確率が高い系統樹はFig2と同じ形である．

5 DISCUSSION

5.1 ワニの進化について

最初に近隣結合法と最尤推定の近似法を用いて系統樹を推定した結果では、ワニが、鳥類やハ虫類や両性類など(以後蜥型類とする)よりもホ乳類に近いという結果が出て、ワニの形態的特徴から見て問題がありそうだった。

しかし、この最尤推定の近似法というのは、もとの系統樹から、少しだけ系統樹を変化させてそれにより尤度が大きくなるような系統樹を信頼できる系統樹とする、という作業を何度も繰り返すことで、最も尤度が高い系統樹を推定するので、大きく系統樹を進化させたほうが尤度が高くても、それを得ることができない。なので、もとの系統樹を変化させたらまた違う結果が出ることがある。

よって、ホ乳類と蜥型類に関して別々に最尤推定の近似法で推定した系統樹を合成して、それでできた系統樹をもとにして最尤推定の近似法をしたところ、ワニが蜥型類側に88％の確率で入る計算結果を得られ、ワニが哺乳類に近い、という発見には結びつきそうにないことがわかった。

5.2 ホ乳類に関して

ホ乳類の進化系統樹は、最尤法のブーストラップ確率も低い部分が多く、対数尤度を求めて推定しても、棄却できない系統樹がたくさんあった。現在信頼されている系統樹でも、ホ乳類の部分は多くの動物が並列表記になっていたので、確率の低い部分についてはあまり言及しないことにした。

5.3 蜥型類の進化

5.3.1 蜥型類について再調査

蜥型類の中に、レアとイグアナと亀とワニが揃った系統樹を得たとき、データを DL した HP で見た系統樹と比べて、亀とイグアナの位置が逆になっていたので、蜥型類の進化に注目して新たに生物を選び直して調査することにした。アウトグループの生物は変えずに、ホ乳類のデータを大幅に減らして、蜥型類の亀とイグアナに近い種の生物を多くとることにした。

5.3.2 新たに得られた蜥型類の進化系統樹

取り直した生物のデータを使って、近隣結合法とそれをもとにした最尤推定の近似法とを行なった。その結果から、信頼できそうな部分系統樹をひとまとめの生物群として扱うことで、6 つの生物群の要素からなる系統樹を作ることができた。その 6 つの要素の対数尤度を求めることで、信頼できる系統樹を推定したところ、Fig5 のような結果が得られた。最も対数尤度が高い系統樹のみが信頼できれば、新たな発見をしたということになる。

5.3.3 蜥型類の進化系統樹の現状

TheTreeOfLife という HP で蜥型類の系統樹を見たところ、亀が蜥型類のどの位置にあるのかについては、?マークがついていて、不確定であることがわかった。さらに詳しく調べると、蜥型類については昔から様々な説が唱えられており、今最も信頼されている系統樹は Fig5 の 2 位の系統樹と一緒の形になっていた。1 位の系統樹は au 値が 0.932 と大きかったが、2 位の系統樹も対数尤度が 0.266 で棄却できないので、新たな発見をすることはできなかった。

6 Conclusion

今回の調査では生物の進化について新たな発見をすることはできなかったが、遺伝子情報をもとにしてこのように統計的手法を用いることで、確かに

信頼できる生物の進化系統樹を推定できるであろうことが確認できた。最尤法がもう少し高速にできたり、遺伝子情報をもとにした新たな推定方法を作ることができれば、より確かな系統樹を得ることができそうな分野であり、研究の余地がある面白い分野なのだろうと思えた。

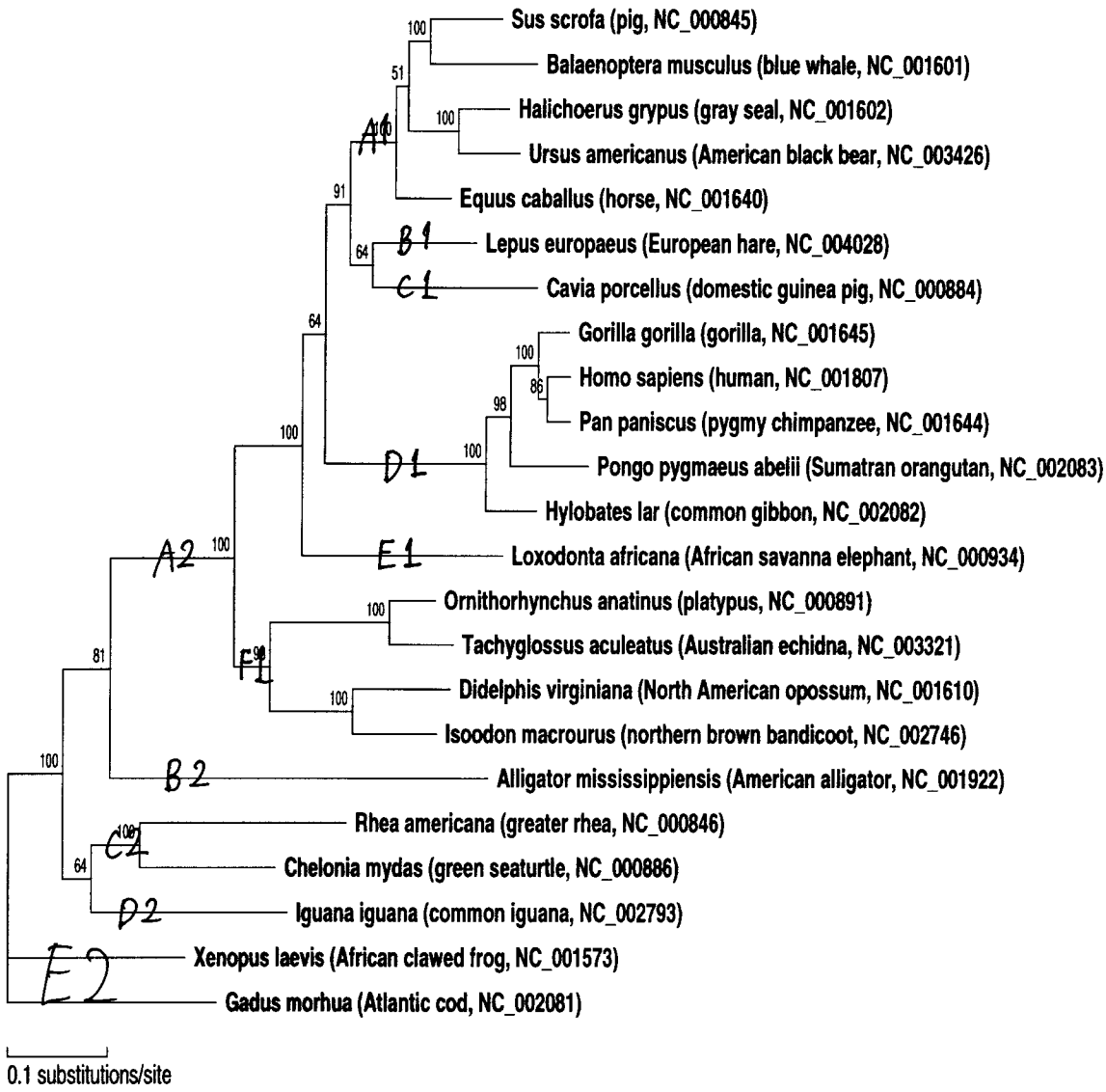

Sus scrofa (pig, NC_000845)
Balaenoptera musculus (blue whale, NC_001601)
Halichoerus grypus (gray seal, NC_001602)
Ursus americanus (American black bear, NC_003426)
Equus caballus (horse, NC_001640)
Lepus europaeus (European hare, NC_004028)
Cavia porcellus (domestic guinea pig, NC_000884)
Gorilla gorilla (gorilla, NC_001645)
Homo sapiens (human, NC_001807)
Pan paniscus (pygmy chimpanzee, NC_001644)
Pongo pygmaeus abelii (Sumatran orangutan, NC_002083)
Hylobates lar (common gibbon, NC_002082)
Loxodonta africana (African savanna elephant, NC_000934)
Ornithorhynchus anatinus (platypus, NC_000891)
Tachyglossus aculeatus (Australian echidna, NC_003321)
Didelphis virginiana (North American opossum, NC_001610)
Isoodon macrourus (northern brown bandicoot, NC_002746)
Alligator mississippiensis (American alligator, NC_001922)
Rhea americana (greater rhea, NC_000846)
Chelonia mydas (green seaturtle, NC_000886)
Iguana iguana (common iguana, NC_002793)
Xenopus laevis (African clawed frog, NC_001573)
Gadus morhua (Atlantic cod, NC_002081)
A1
B1
C1
D1
E1
A2
F1
B2
C2
D2
E2
0.1 substitutions/site

Fig 1

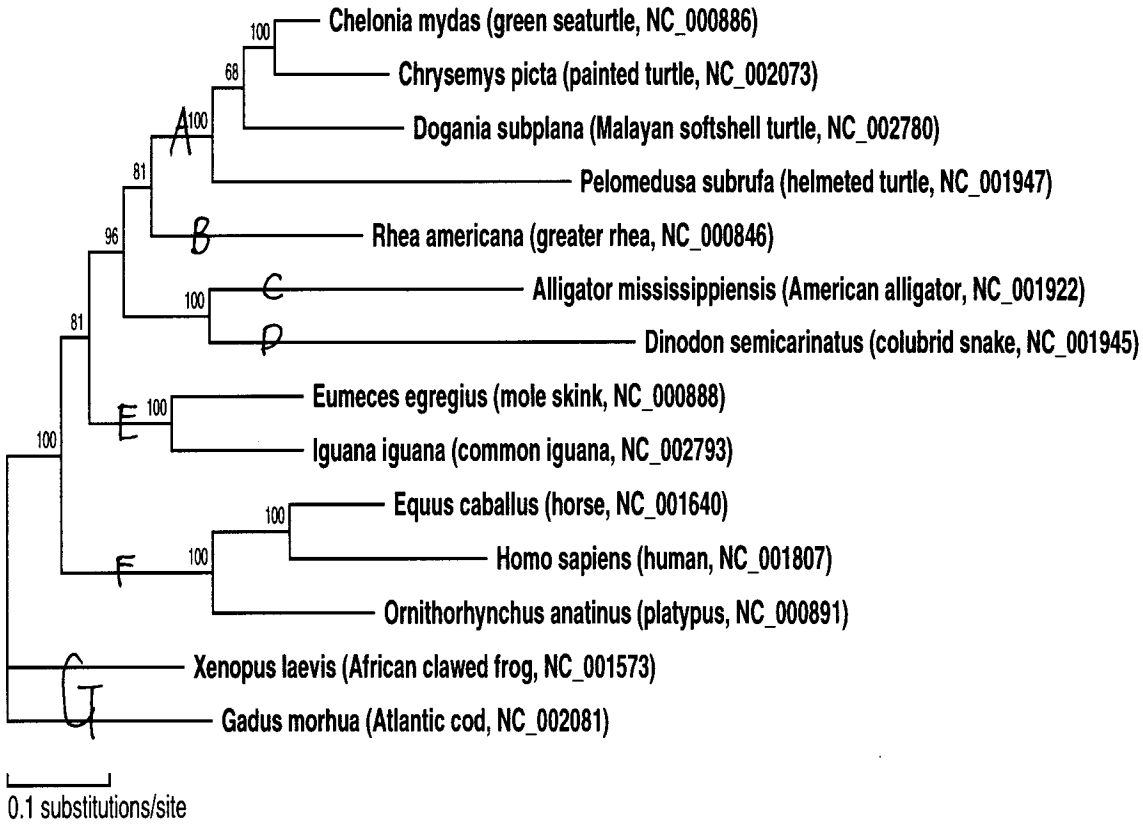

Chelonia mydas (green seaturtle, NC_000886)
Chrysemys picta (painted turtle, NC_002073)
Dogania subplana (Malayan softshell turtle, NC_002780)
Pelomedusa subrufa (helmeted turtle, NC_001947)
Rhea americana (greater rhea, NC_000846)
Alligator mississippiensis (American alligator, NC_001922)
Dinodon semicarinatus (colubrid snake, NC_001945)
Eumeces egregius (mole skink, NC_000888)
Iguana iguana (common iguana, NC_002793)
Equus caballus (horse, NC_001640)
Homo sapiens (human, NC_001807)
Ornithorhynchus anatinus (platypus, NC_000891)
Xenopus laevis (African clawed frog, NC_001573)
Gadus morhua (Atlantic cod, NC_002081)
100
68
100
81
96
100
81
100
100
100
100
A
B
C
D
E
F
G
0.1 substitutions/site

Fig 2

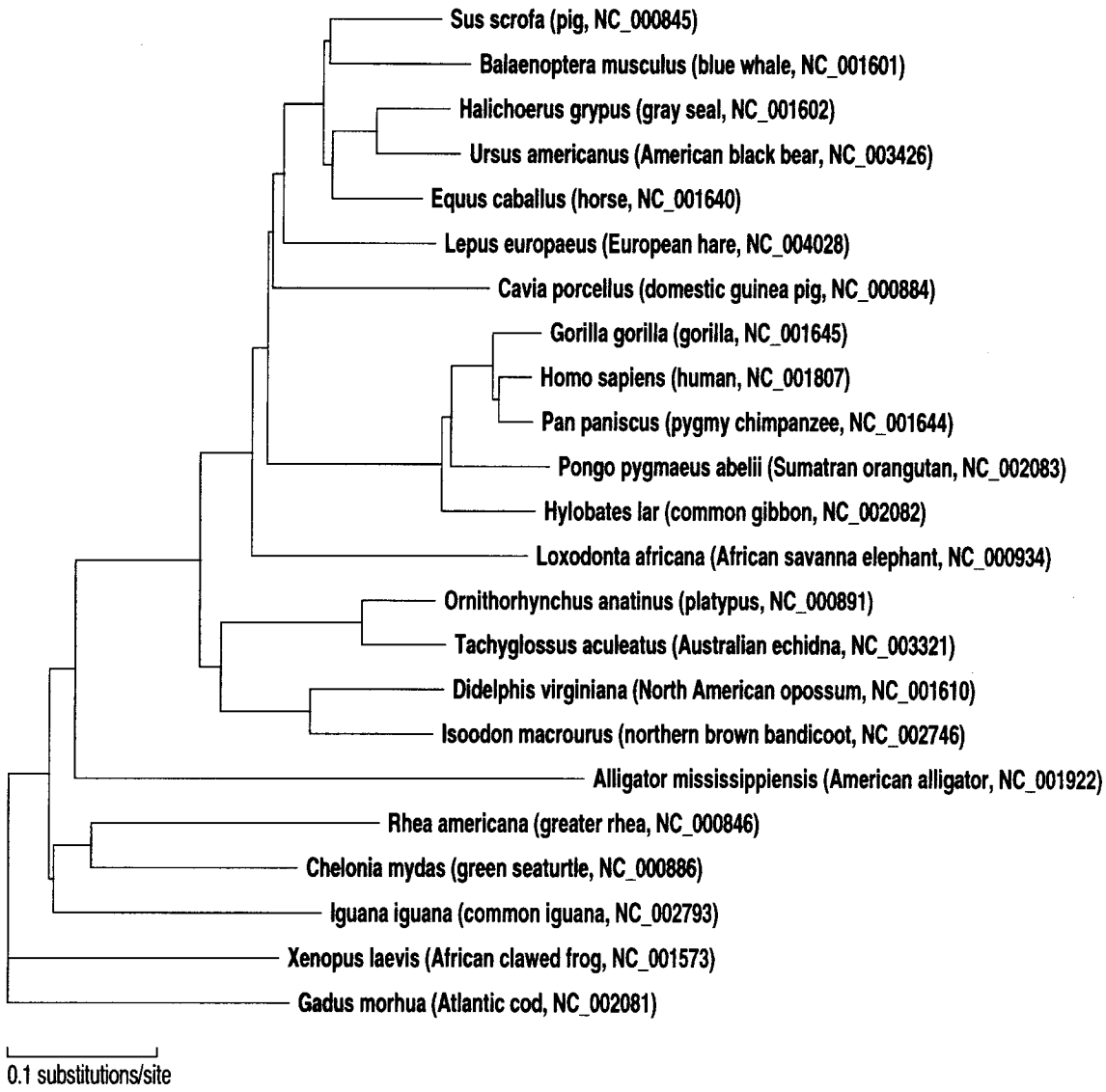

Sus scrofa (pig, NC_000845)
Balaenoptera musculus (blue whale, NC_001601)
Halichoerus grypus (gray seal, NC_001602)
Ursus americanus (American black bear, NC_003426)
Equus caballus (horse, NC_001640)
Lepus europaeus (European hare, NC_004028)
Cavia porcellus (domestic guinea pig, NC_000884)
Gorilla gorilla (gorilla, NC_001645)
Homo sapiens (human, NC_001807)
Pan paniscus (pygmy chimpanzee, NC_001644)
Pongo pygmaeus abelii (Sumatran orangutan, NC_002083)
Hylobates lar (common gibbon, NC_002082)
Loxodonta africana (African savanna elephant, NC_000934)
Ornithorhynchus anatinus (platypus, NC_000891)
Tachyglossus aculeatus (Australian echidna, NC_003321)
Didelphis virginiana (North American opossum, NC_001610)
Isoodon macrourus (northern brown bandicoot, NC_002746)
Alligator mississippiensis (American alligator, NC_001922)
Rhea americana (greater rhea, NC_000846)
Chelonia mydas (green seaturtle, NC_000886)
Iguana iguana (common iguana, NC_002793)
Xenopus laevis (African clawed frog, NC_001573)
Gadus morhua (Atlantic cod, NC_002081)
0.1 substitutions/site

Fig 3

njdist 2.3b3 11/06/02 14 OTUs 3481 sites mtREV24-F ND1 ND2 COX1 COX2 ATP8 ATP6 COX3 ND3 ND4L ND4 ND5 CYTB

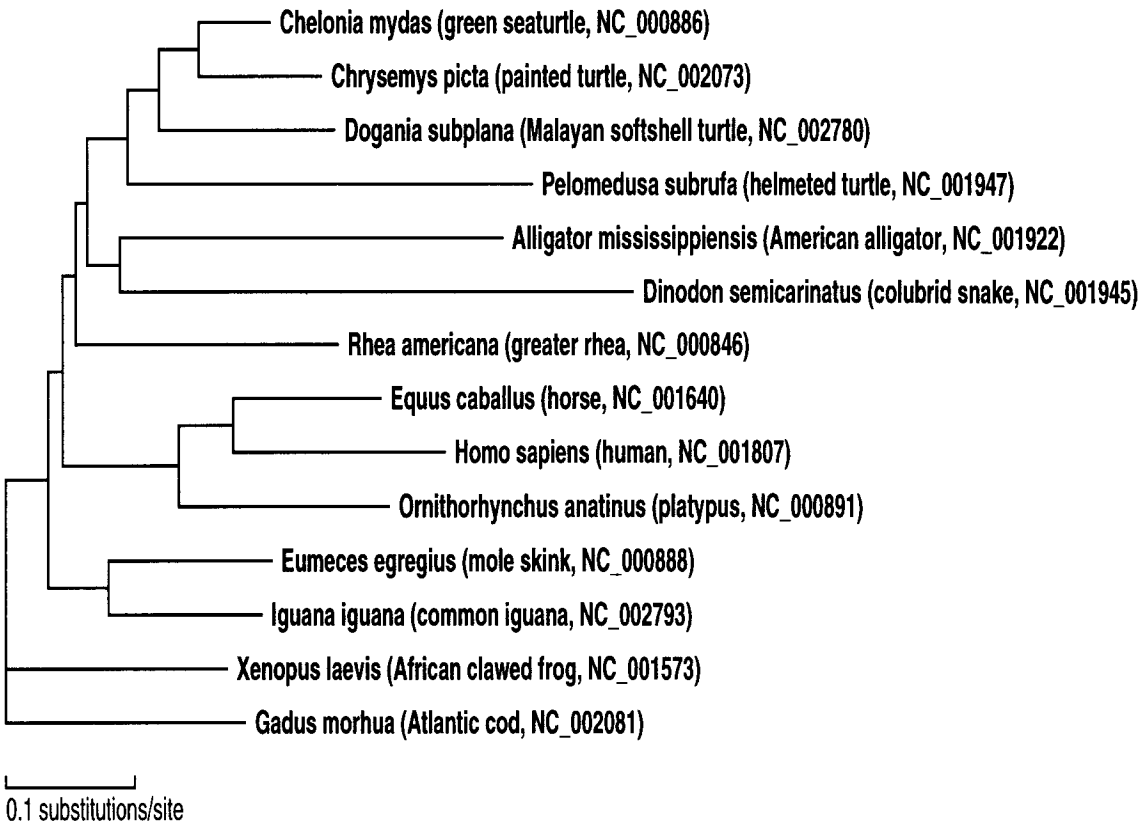
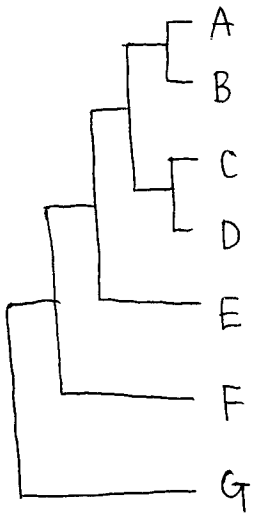
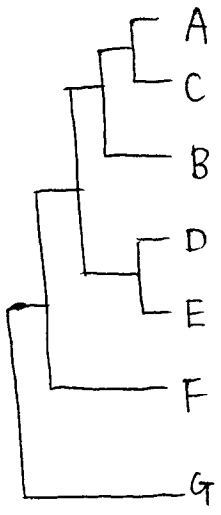


Fig 4

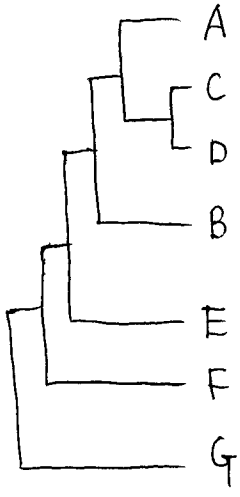
TREE



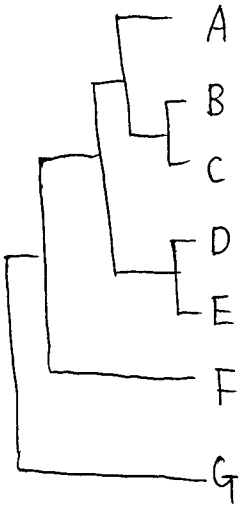
1位 au : 0.932



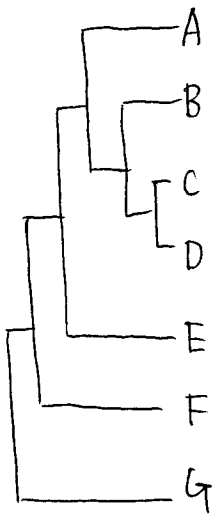
2位 au: 0.266



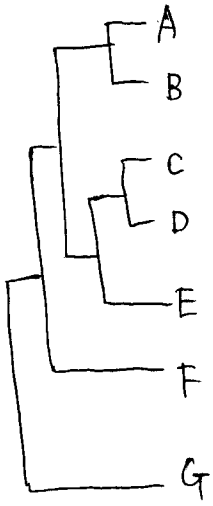
3位 au: 0.242



4位 au: 0.210



5位 au: 0.175



6位 au: 0.075

EDGE

1位	A B	(au: 0.882)	7位	D E	(0.191)
2	A B C D	(0.873)	8	A B C	(0.181)
3	C D	(0.820)	9	B C D	(0.174)
4	A C	(0.255)	10	C D E	(0.074)
5	A C D	(0.241)			
6	B C	(0.201)			

Fig. 5