モンテカルロ法による分岐図作成のための最適化法

東京農工大学大学院
連合農学研究科
松田　朋子

次世代シークエンサーによる遺伝子解析
大量のデータをどう扱うか？

サンガーシークエンサー

次世代シークエンサー

リード長　約 1,000 bp
１ランあたりの時間　　２時間
１ランあたりのデータ量
約　0.016 Mbp

リード長　 100 bp
１ランあたりの時間　　約4日間
１ランあたりのデータ量
約　95 ,000 Mb

解読量が大幅に増加　⇒　データ解析が困難

目的
どの遺伝子を使うのか？

- 従来，分岐図は，数個の遺伝子（ミトコンドリアDNAの一部の配列など）をもとに作成していた
- 近年のゲノム解析技術の進歩にともない，ゲノム全体にわたる大量の遺伝子情報を容易に得られるようになった
- 大量の遺伝子情報のうち，どの遺伝子を使って分岐図を作成するのかを検討する必要が出てきた
- 今回は，大量のデータから無作為にサンプルを抽出する方法（モンテカルロ法）を用いて，無数の遺伝子の組合せを作る方法を確立することを目的とした

方法

1. 組合せ生成
   - 447種類の遺伝子の組合せ数を計算して，組合せのパターンを生成
   - 447種類の遺伝子のうち1～4遺伝子を除く組合せの数は“1656133808 “
2. 乱数と数列の作成
   - “ 1656133808 “ から一様乱数を発生させる
   - 乱数に対応する組合せ数列を作成
3. 分岐図作成
   - RAxMLによる分岐図作成
4. 分岐パターン抽出
   - 系統樹を推定し，分岐パターンを抽出

結果①：組合せ生成

447個の遺伝子から任意の個数の遺伝子を選択する組合せの数を計算

- > CHOOSE(447,1)
  [1] 447
- > CHOOSE(447,1)+CHOOSE(447,2)
  [1] 100128
- >CHOOSE(447,1)+CHOOSE(447,2)+CHOOSE(447,3)
  [1] 14886143
- >CHOOSE(447,1)+CHOOSE(447,2)+CHOOSE(447,3)+CHOOSE(447,4)
  [1] 1656133808 <- 本研究で使用した乱数の数. 選択した1~4個の遺伝子を解析から除く

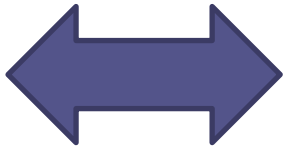
結果②：乱数の発生と対応する組合せ数列の選択

1. 一様乱数
   RUN1 <- ROUND(RUNIF(100000, MIN = 1, MAX = 1656133808))

2. 447個の遺伝子から4個までを選択する1656133808個の組合せ数列を産生

3. 乱数と数列を対応させる

発生させた乱数
81571247
365913044
1023396239
644416555
・
・



乱数に対応する数列

545682	420	121	150	372
276325	222	29	176	229
9585546	17	160	302	413
81571247	20	199	232	430

・
・
・

遺伝子IDに対応

4. 447個のうち乱数に対応する遺伝子を除き，残りのすべての遺伝子をつかって分岐図を作成

結果③：RAxMLによる分岐図作成

- 分岐図作成ソフトRAxMLを実行
- RAxML の出力結果から数値や記号などを取り去り，分岐図パターンをシンボリックデータにする

RAxML の出力結果

(((X:0.0053140481372246041 1,B:0.0058268092980178331 4):0.0259712662504384093 9,(M:0.0364532495599415445 9,((I:0.0213012216007929973 4,G:0.0314479076574554206 0):0.0042415128186705456 5,U:0.0294868676965653678 2):0.0204036004694002175 2):0.0152177499489961100 3):0.0084080621906077023 7,((((Z:0.3434476718995415622 8,(T:0.1426261395456087932 6,L:0.0961102430657040651 7):0.1248272718875478165 5):0.1773867938945919986 4,(((Q:0.0922090363633366744 1,C:0.0784974040296460562 3):0.0216221371004468726 5,(((((E:0.0220962201887033929 4,K:0.0289328311909968147 2):0.0089715453504747855 2,P:0.0852616742679427608 3):0.0139319394947335466 2,(F:0.0131835663044479935 9,(N:0.0195249052320230279 1,&:0.0169312730877942511 9):0.0081360492881540000 3):0.0781798885546699240 4):0.0226343231508473220 8,(R:0.0459044576751964084 1,H:0.0463731538899979714 4):0.0386531867966414532 9):0.0078919137381470525 6,(V:0.0328461191799640991 9,D:0.0225021056831853257 0):0.1792303279841169216 8):0.0100104948709819272 0):0.1032573418974204876 3,(Y:0.2008624935488638185 7,(J:0.1832175297537883274 0,#:0.2096798539032603270 2):0.1299285832980961452 6):0.0188052384532221769 6):0.0385741459174890263 8):0.0662010103190122239 9,(O:0.0385464852009356068 2,S:0.0355270492745269894 6):0.1270685517072865922 1):0.0563083003690214994 9,W:0.1393362962639454749 6):0.0715778040946137700 3,A:0.0378202172015906640 2):0.0;

⇒簡素化した分岐図パターン

((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)

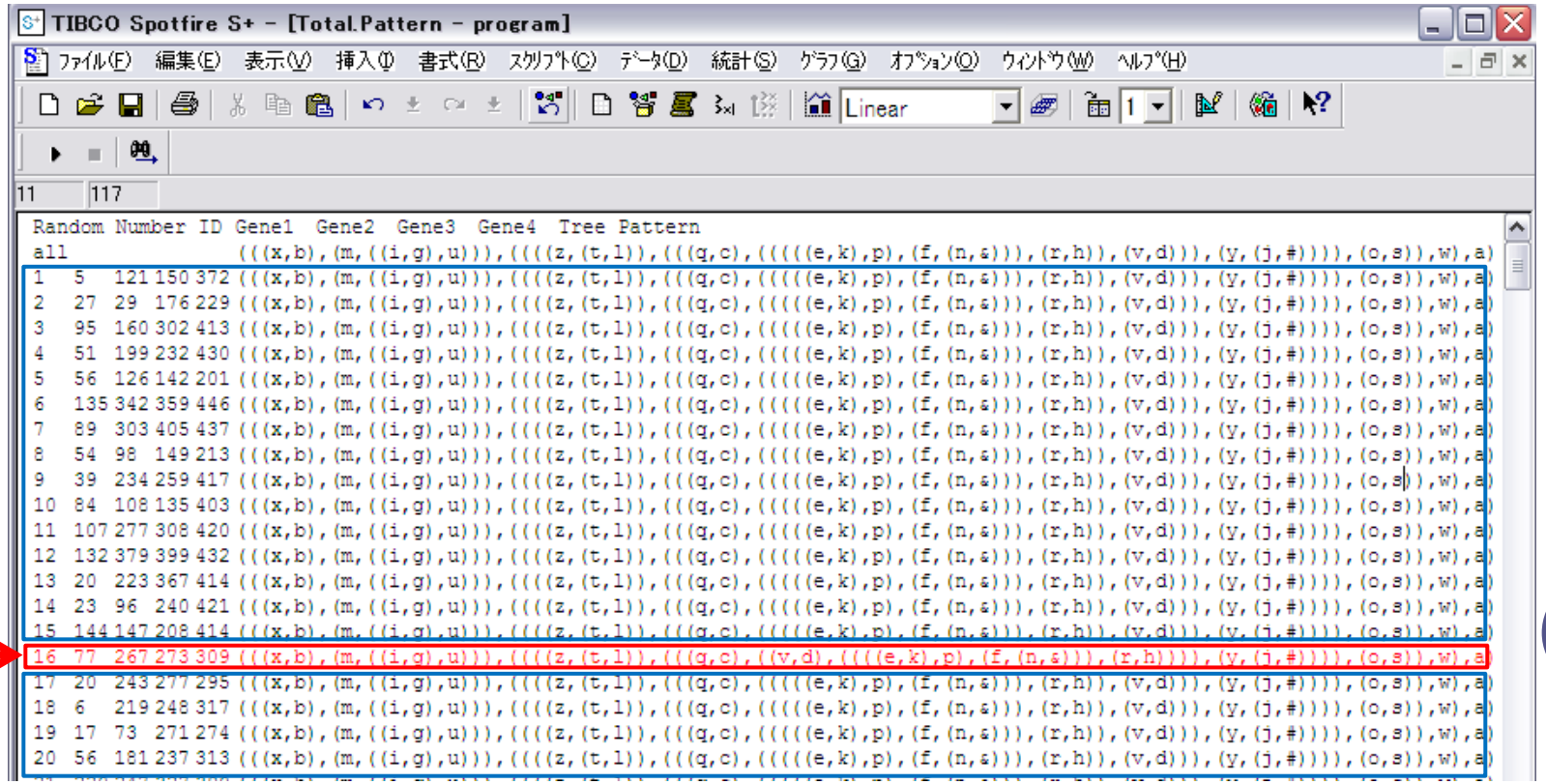
結果④：分岐図パターンの頻度

185/200 試行 で得られた樹形

(((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)

15/200 試行で得られた樹形

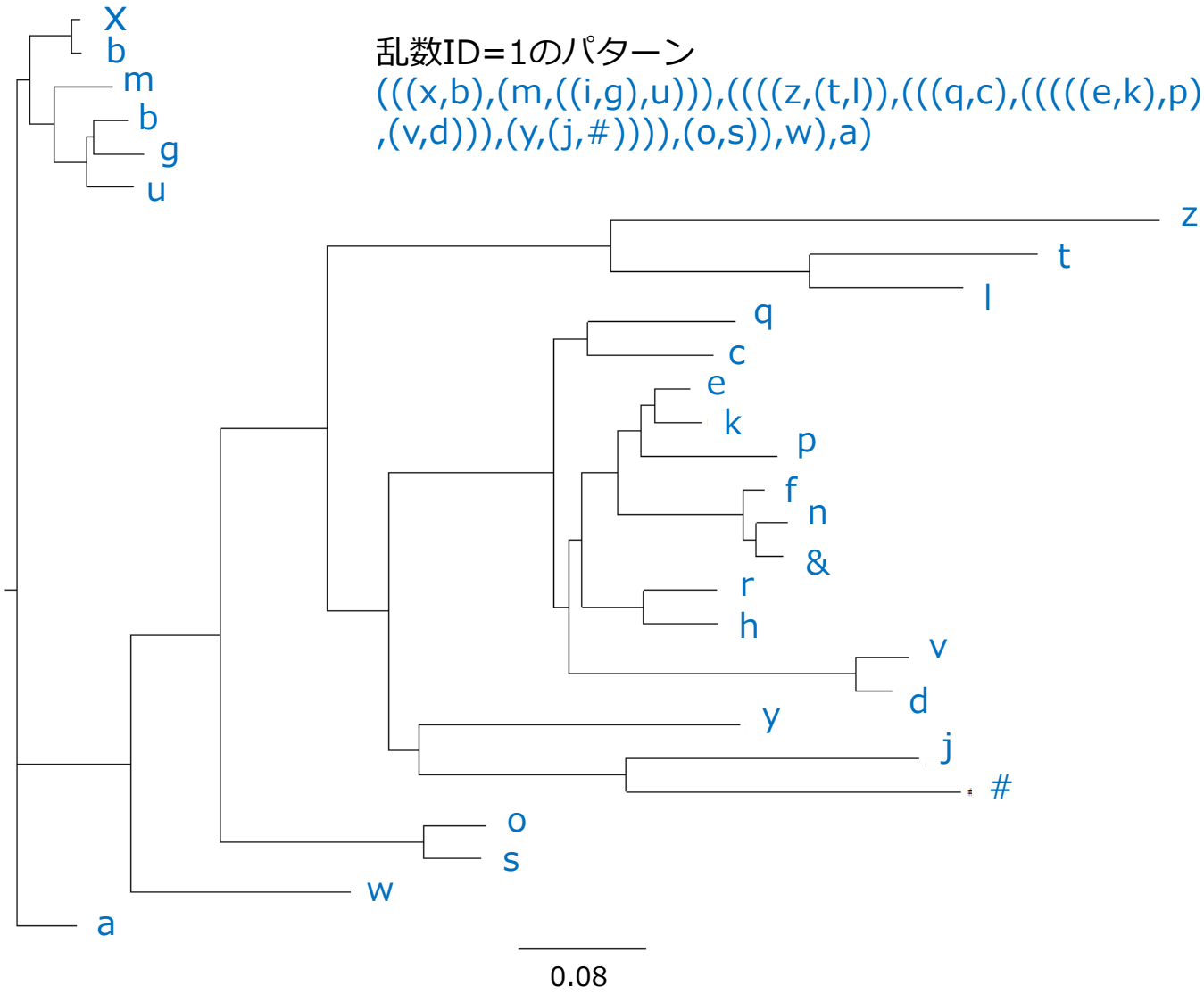
(((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((v,d),((((e,k),p),(f,(n,&))),(r,h)))),(y,(j,#)))),(o,s)),w),a)



結果⑤最適化された分岐図パターン
（185/200 回の試行で得られたパターン）

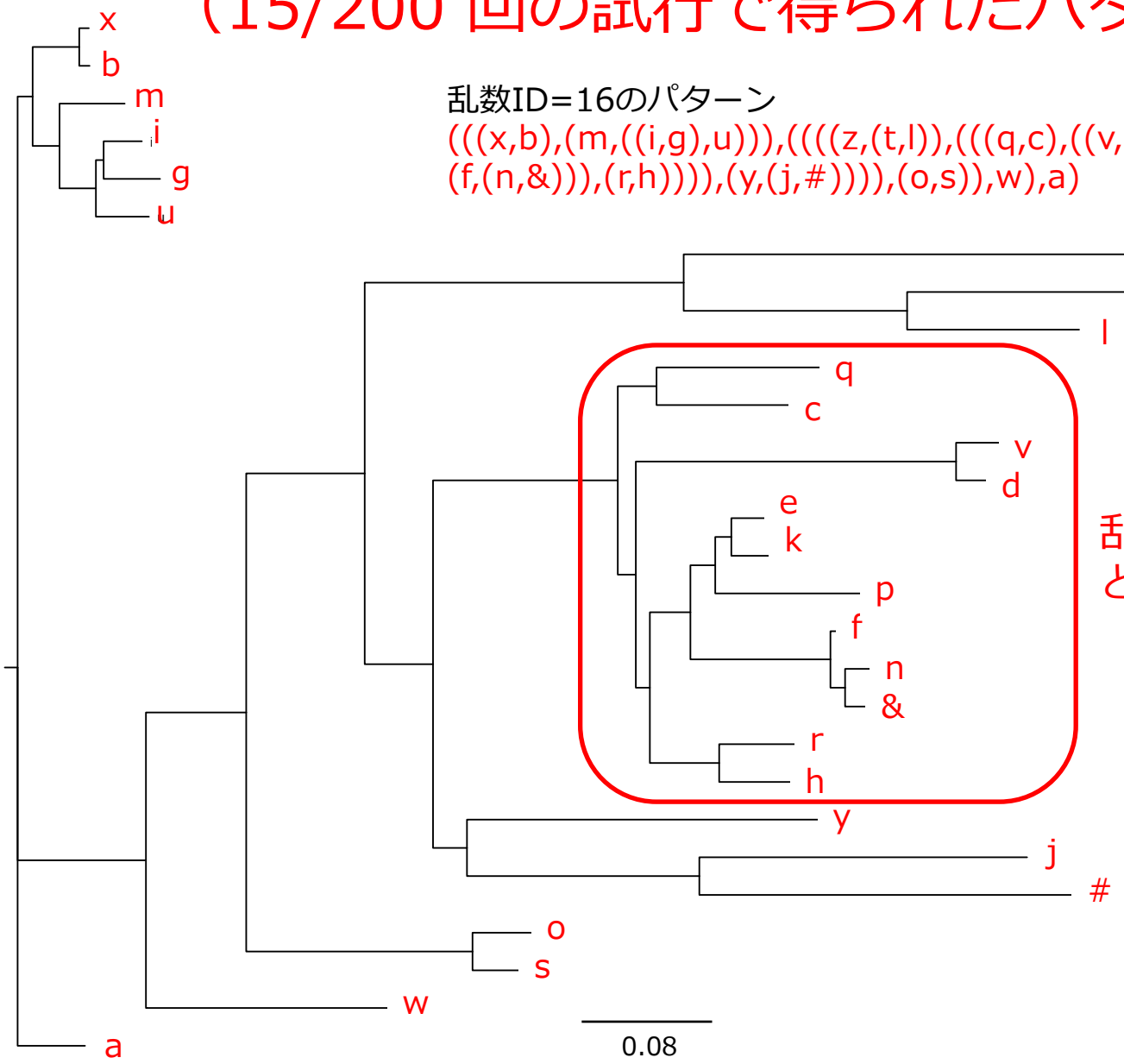
乱数ID=1のパターン

(((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)



結果⑥最適化された分岐図とは異なるパターン
（15/200 回の試行で得られたパターン）

乱数ID=16のパターン
(((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)



乱数1のパターンと異なる

まとめ

1. 大量のデータから無作為にサンプルを抽出する方法（モンテカルロ法）を用いて，無数の遺伝子の組合せを作る方法を確立した

2.　200回の試行のうち，185回の試行で完全に一致するパターンが得られた．一方，残り15回の試行では，異なるパターンが得られた．

3.　今後は，より多くの試行を実施して，異なるパターン（　15/200　回の試行）が得られた時に除かれた遺伝子（=185/200回の試行で得られたパターンを得るために必要な遺伝子であると考えられる）をリストアップする．リストアップした遺伝子の中には，分岐図作成に強く関与している遺伝子が含まれる可能性が高いと考えられる．