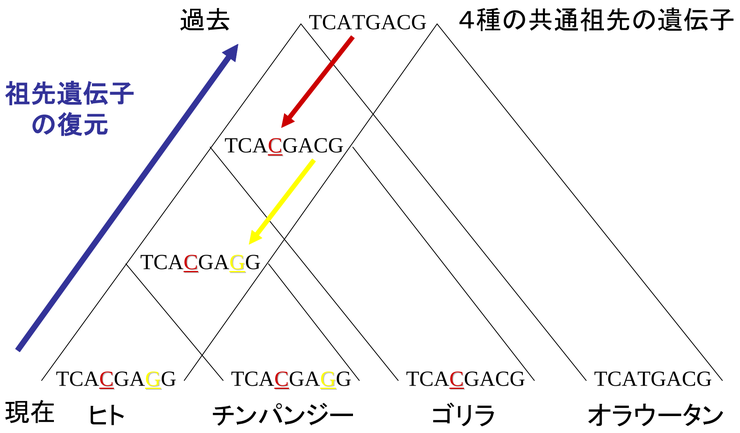
分子進化のモデリングと分子系統樹の推定

2014年6月13日　統計数理研究所　オープンハウス

足立　淳　　データ科学研究系　准教授

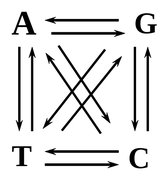
分子系統学とは

遺伝子の変異差から種の系統を推定



配列の違いは共通祖先から進化してきた結果生じたものであり、進化の歴史を反映している。

→ 最尤法による系統樹推定

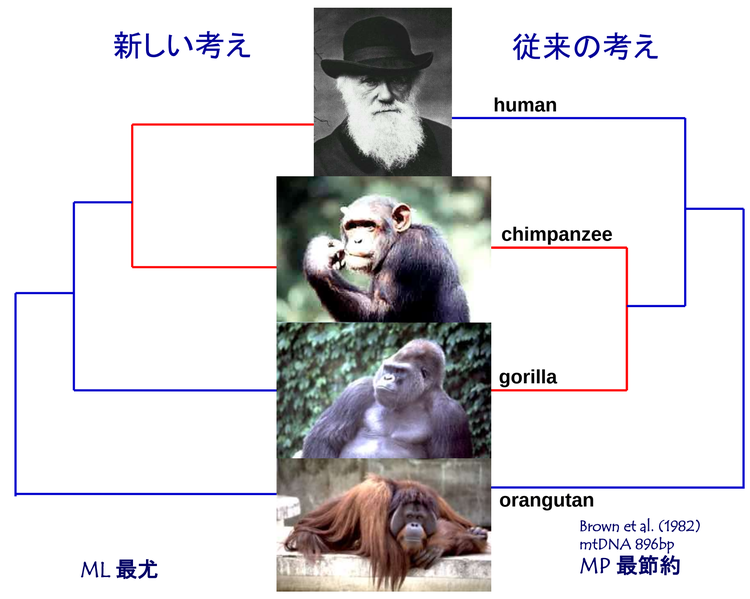


塩基置換の確率モデル

1. Human　ヒト
2. Chimpanzee　チンパンジー
3. Gorilla　ゴリラ
4. Orangutan　オランウータン

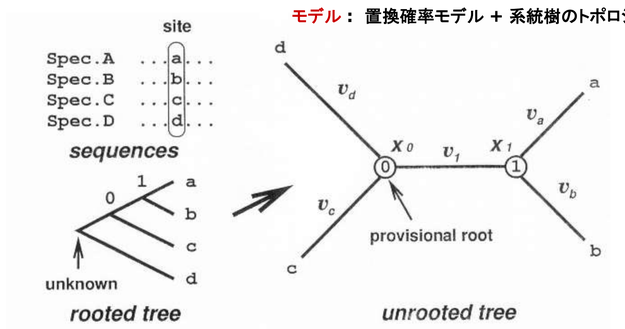
確率過程
- 突然変異
- 変異遺伝子の集団への固定

```
1 CTAGGCTATATACAACTACGCAAAGGCCCCAACGTTGTAGGCCCCTAC
2 CTAGGCTACATACAACTACGCAAAGGTCCCAACATTGTAGGTCCTTAC
3 TTAGGCTATATACAACTACGTAAAGGCCCCAACGTCGTAGGCCCCTAC
4 CTAGGCTATACACAACTACGCAAGGGACCTAACATCGTAGGCCCCTGC
```



最尤法と尤度(Likelihood)　$L = P(\text{data}|\text{model})$

モデル： 置換確率モデル ＋ 系統樹のトポロジー



尤度は、与えられたモデルのもとで、データが実現する確率。
最尤法ではこれを最大化するモデルを選択。
モデルに含まれるパラメータも最尤推定。

Likelihood of a site

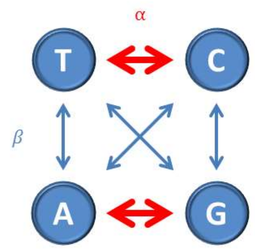
$$L = \sum_{x_0} \pi_{x_0} [P_{x_0 c}(v_c) \times P_{x_0 d}(v_d) \times \sum_{x_1} P_{x_0 x_1}(v_1) P_{x_1 a}(v_a) P_{x_1 b}(v_b)]$$

置換モデル

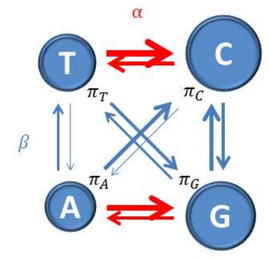
塩基置換やアミノ酸置換のマルコフモデル
Transition probability matrix P(t) during time t
$\mathbf{P}(t) = e^{\mathbf{Q}t}$

- DNAモデル（塩基置換）
  - A ⇔ G,　T ⇔ C　トランジション
  - (A,G) ⇔ (T,C)　トランスバージョン
- タンパクモデル（アミノ酸置換）
  - 経験的に得られた 20×20 の置換配列
- コドンモデル
  - アミノ酸置換 と 同義塩基置換 の組み合わせ
  - 深い分岐の系統と浅い分岐の系統を同時推定

塩基置換モデル



Transition / Transversion model
K80
Kimura's 2 parameter (1980)



Ts / Tv + Proportional model
HKY85
Hasegawa, Kishino & Yano 1985

コドンモデル　CTA へ9通りの置換

code	Amino acid	code	Amino Acid	code	Amino acid	code	Amino acid	
TTT	F Phe	TCT	S Ser	TAT	Y Tyr	TGT	C Cys	T
TTC		TCC		TAC		TGC		C
TTA	L Leu	TCA		TAA	- ---	TGA	- ---	A
TTG		TCG		TAG	- ---	TGG	W Trp	G
CTT		CCT	P Pro	CAT	H His	CGT	R Arg	T
CTC		CCC		CAC		CGC		C
CTA		CCA		CAA	Q Gln	CGA		A
CTG		CCG		CAG		CGG		G
ATT	I Ile	ACT	T Thr	AAT	N Asn	AGT	S Ser	T
ATC		ACC		AAC		AGC		C
ATA		ACA		AAA	K Lys	AGA	R Arg	A
ATG	M Met	ACG		AAG		AGG		G
GTT	V Val	GCT	A Ala	GAT	D Asp	GGT	G Gly	T
GTC		GCC		GAC		GGC		C
GTA		GCA		GAA	E Glu	GGA		A
GTG		GCG		GAG		GGG		G
T		C		A		G		

1st codon position　3rd codon position　2nd codon position

同義置換（塩基）　非同義置換（アミノ酸）

非同義置換（アミノ酸）

EF1α のアミノ酸配列：動物、菌類、植物、原生生物、細菌で共通の配列が見られ

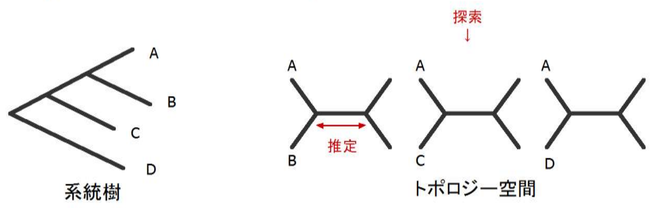
```
CONSENSUS  STTTGHLIYK CGGIDKRTIE KFEKEAAE.G KGSFKYAWVL DKLKAERERG ITIDIALWKF ET.KY.VT.I DAPGHRDFIK
Homo sapiens  ..........  ..........  ........M.  ..........  ..........  ....S.....  .S..Y..I.  ..........
Gallus gallus  ..........  ..........  ........M.  ..........  ..........  ....S.....  .S..Y..I.  ..........
Xenopus laevis  ..........  ..........  ........M.  ..........  ..........  ....S.....  .S..Y..I.  ..........
Danio rerio  ..........  ..........  ........M.  ..........  ..........  ....S.....  .S..Y..I.  ..........
Apis mellifera  ..........  ..........  .....Q.M.  ..........  ..........  ..........  .A..Y..I.  ..........
Bombyx mori  ..........  ..........  .....Q.M.  ..........  ..........  ..........  .S..Y..I.  ..........
Onchocerca  ..........  ..........  .....Q.M.  ..........  ..........  .Q........  .P..YI.I.  ..........
Saccharomyces  ..........  ..........  .....L..  ..........  ..........  ..........  .P..Q..V.  ..........
Ashbya gossypi  ..........  ..........  .....L..  ..........  ..........  ..........  .P..H..V.  ..........
Candida albica  ..........  ..........  .....L..  ..........  ..........  ..........  .P..H..V.  ..........
Trichoderma re  .....Q....  ..........  .....L..  ..........  ..........  ..........  .P..Y..V.  ..........
Podospora anse  ..........  ..........  .....L..  ..........  ..........  ..........  .P..Y..V.  ..........
Puccinia grami  ..........  ..........  .....L..  ..........  ..........  ..........  .P..Y..V.  ..........
Absidia glauca  ..........  ..........  .....L..  ..........  ..........  ..........  .P..H..V.  ..........
Arabidopsis th  ..........  L.....V..  R.....MN  .R........  ..........  ....S.....  .T..YC.V.  ..........
Glycine max  ..........  L.....V..  R.....MN  .R........  ..........  ..........  .T..YC.V.  ..........
Hordeum vulgar  ..........  L.....V..  S.....MN  .R........  ..........  ..........  .T..SC.V.  ..........
Triticum aesti  ..........  L.....V..  R.....MN  .R........  ..........  ..........  .T..YC.V.  ..........
Trichomonas te  ..........  ..L...KLA  AM...EQL.  .S.....F.M  .S........  ....S.....  GQ..SF.I.  ..........
Giardia lambli  .L........  .....Q..D  EY..R.T.M.  ..........  .Q.D......  ..N.......  .K..I..I.  ..........
Hexamitidae  .L........  .....Q..LD  EY..R.N.M.  ..........  .Q.D......  ..N.......  .K.FT..I.  ..........
Hexamita infla  .L........  .....Q..L.  DY..K.N.I.  ..........  .Q.D......  ..N.......  .K.FI..I.  ..........
Glugea plecogl  ..V.N.AFQ  L.A...A.KMD  .LK...E.R.  R.T.S..Y.M  .MSA......  ...TTS.M.L  ..S.HMLNV.  .C..G....
Sulfolobus aci  .LI.R.LMD  R.F..EK.VK  EA.EA.KKL.  .D.E...FLM  .R..E.....  V..NLSFMR.  ..R..FF.V.  .......V.
Halobacterium  .LV.R.L.E  T.SVPEHV..  QHKE..E.K.  ..G.EF.Y.M  .N.AE.....  V.....HQE.  S.DT.DF.IV  .G.....V.
Methanococcus  ..V.R.LLD  G.A..POL.V  RLR...E.K.  .AG.EF.Y.M  .G..E.....  V..V.HK..  P.A..E..IV  .G.......
                10        20        30        40        50        60        70        80
```

アミノ酸置換モデル　mtREV

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A		31	135	69	328	41	63	727	84	595	165	1	772	35	321	2229	2078	1	41	818
R	31		135	1	1	1102	1	152	455	1	71	985	1	34	178	38	34	146	1	51
N	135	135		4527	30	731	503	314	2656	132	151	2712	266	117	508	2607	1271	67	893	76
D	69	1	4527		1	261	3495	334	828	88	4	175	1	35	1	319	98	63	74	1
C	328	1	30	1		242	1	214	723	256	247	1	1	547	101	1788	1046	222	897	1
Q	41	1102	731	261	242		1798	83	3208	72	281	2631	349	160	801	447	524	1	260	78
E	63	1	503	3495	1	1798		191	306	1	5	1630	1	1	54	286	138	1	115	137
G	727	152	314	334	214	83	191		1	40	10	89	5	1	1	651	73	49	1	20
H	84	455	2656	828	723	3208	306	1		43	62	470	13	177	273	341	295	47	3171	1
I	595	1	132	88	256	72	1	40	43		1875	65	2656	441	117	280	1805	1	200	6030
L	165	71	151	4	247	281	5	10	62	1875		37	2938	1112	184	388	717	190	231	465
K	1	985	2712	175	1	2631	1630	89	470	65	37		440	52	293	765	808	183	233	6
M	772	1	266	1	1	349	1	5	13	2656	2938	440		546	89	714	2547	209	297	1745
F	35	34	117	35	547	160	1	1	177	441	1112	52	546		93	310	267	57	2534	34
P	321	178	508	1	101	801	54	1	273	117	184	293	89	93		720	732	33	102	48
S	2229	38	2607	319	1788	447	286	651	341	280	388	765	714	310	720		3426	148	218	1
T	2078	34	1271	98	1046	524	138	73	295	1805	717	808	2547	267	732	3426		125	158	1172
W	1	146	67	63	222	1	1	49	47	1	190	183	209	57	33	148	125		151	35
Y	41	1	893	74	897	260	115	1	3171	200	231	233	297	2534	102	218	158	151		36
V	818	51	76	1	1	78	137	20	1	6030	465	6	1745	34	48	1	1172	35	36	

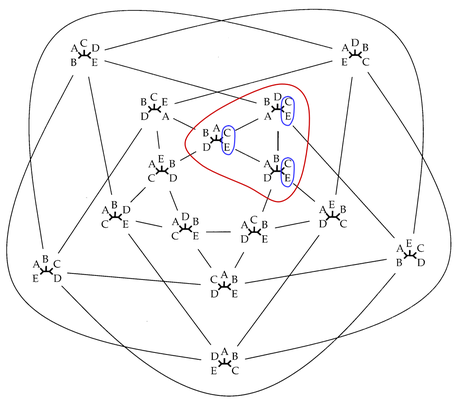
系統樹 ＝ トポロジー ＋ 距離

- トポロジー　＝　枝分かれの順番
  トポロジー空間から探索
- 距離　＝　枝の長さ（推定置換数）
  置換確率モデルから計算



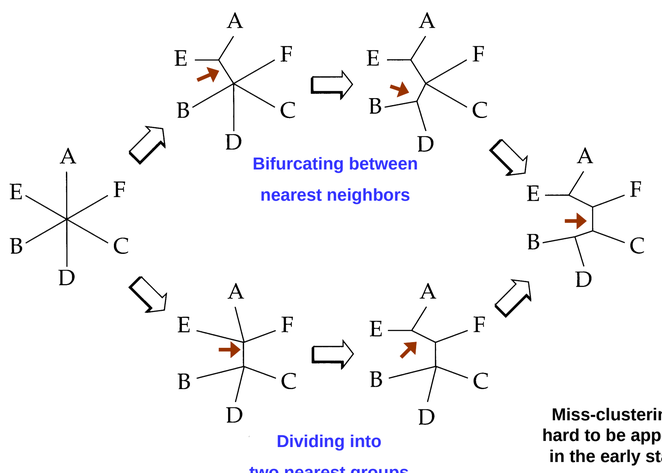
トポロジー空間と探索法

最近隣木間の関係

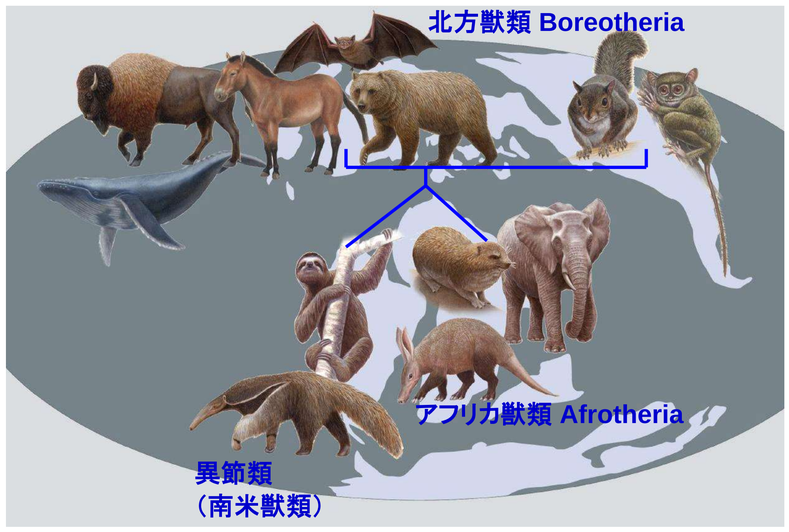


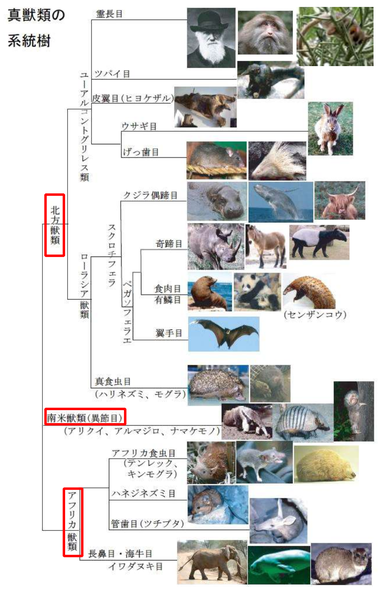
The space of all 15 possible unrooted trees with 5 tips

New star decomposition



1億年前の大陸配置と真獣類の進化





遺伝子系統×ゲノム構造比較＝進化ダイナミクス