数理解析研究所講究録
第1597巻 2008年 191-197

家系図ネットワークの構造解析

大阪府立大学大学院工学研究科数理工学分野　水口 毅 (Tsuyoshi Mizuguchi) [1]
Department of Mathematical Sciences, Osaka Prefecture University
京都大学大学院情報学研究科数理工学専攻　西村 麻衣子 (Maiko Nishimura)
Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University

1　はじめに ～祖先数のパラドクス

ヒトには父と母がいる．そして，その父母にもそれぞれにも父と母がいるので，二世代前の祖先は4人である．同様に，どの祖先にも父と母がいることから，三世代前の祖先は8人，四世代前は16人とご先祖の数は2の巾で指数関数的に増大していき，ついにはその時代の総人口を越えてしまう…．このパラドクスは，全てのご先祖を別々にカウントしてしまったことに起因していると考えざるを得ない．あなたには重複しているご先祖様がいるのだ（しかもたくさん重複している！）．この絡み合っているに違いないご先祖さまたちの関係を数理的な問題としてきちんと設定するために，親子関係を単純かつ明解に表す方法として，親子関係にある個体同士を線で結んだ図—いわゆる家系図（図1）—を考えてみよう．すると，上記の問題は家系図の構造を調べることに他ならない．この家系図はどのようなネットワーク構造をしているのだろうか？まずは素朴な切口として，(i) あるヒトの家系図はどの程度重複しているか？(ii) 十分過去に遡ったら，その時代に生きているヒトはみなご先祖様か？などの疑問が挙げられよう．

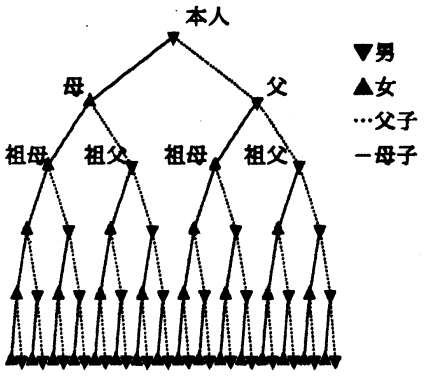

図1. もっとも単純な家系図.

ヒトに限らず両親性 (biparental) な生物ならば文字通り両親が必ずいるので，上記の問題は必ず起こる．家系図や分岐過程での統計的な性質に関しては，様々な先行研究が行われている [1, 2, 3, 4] が，ここでは Derrida らによるものを紹介しよう [3]．彼らは，両親性の仮想的な生物個体集団に対して，4つの仮定

A1. 親は集団内でランダムに対を作り子供を作る

A2. 性差は存在しない

A3. 世代重複はない

A4. 子供の数は平均 m のポアソン分布である

を採用し，仮想的な家系図を構成，その家系図の構造を，理論および数値計算によって解析した．いくつもの興味深い結論が得られているが，その中に，「ある個体から G 世代前の祖先の数を $A(G)$，G 世代前の総個体数を $N(G)$ とすると，十分大きな G に対して「非先祖率」 $S(G) \equiv 1 - A(G)/N(G)$ は，平均子数 m だけで決まる一定値 $S^* < 1$ に漸近する.」という結果がある．この結果は，巧妙な理論および数値計算を元に主張されているが，実際の生物での検証は行われていない．

と言うのも，実際の生物で親子関係が何世代にもわたって保存されている例はなかなか見当たらないからである．ヒトの場合，普通3～4世代もさかのぼれば，記憶があやふやになり，墓石や過去帳を調べなければならない．天皇家，王家，貴族，大名などいわゆる「高貴な」家には家系図が残されていることが多いが，それらは基本的に「男性」の血統であり，「女性」のそれは基本的に無視されていることがほとんどである[2] [3]．

[1] corresponding author: gutchi@ms.osakafu-u.ac.jp
[2] 系図売りという商売もあったことを考えると，信憑性の問題もあるが，ここでは目をつぶるとしよう．
[3] 名字（苗字）も親から受け継がれるものであるが，両親性ではないので，祖先数のパラドクスは起きない．

本稿では，Derrida らによって得られたこの結果が実際の生物でも妥当かどうかを解析する．対象となる生物は競走馬である．競走馬は WEB 上に公開されているデータベースから比較的容易に家系をたどることができ，十数世代にわたる家系図を構築することが可能である．

本論文の構成は以下の通りである．２節では実在する家系図として競走馬のデータの解析を行った．その結果，４つの仮定のうち，いくつかは破られていることが明らかになった．また，破れた仮定を考慮し，新たに世代および先祖率を定義し理論と観測値の比較を行ったところ，両者は必ずしも一致していないという結果が得られた．３節ではまとめと考察および今後の展望を行う．

2　データと解析 ～連綿と続く競走馬の家系

本稿では競走馬—いわゆる「サラブレッド」の家系に着目する．競走馬の成績には，その血統が重要視されており，家系のデータも比較的容易に入手できる．今回は，インターネット上のデータサイト「Thoroughbred Horse Pedigree Query(www.pedigreequery.com)」に登録されている馬を対象とする[4]．

用語の定義：まず注目する一個体を定め，「主個体」と呼ぶ．主個体 α の先祖とは，「主個体の」＋{「父の」もしくは「母の」の０個以上の任意の組合せ}＋{「父」あるいは「母」}で表される個体とする．主個体とその先祖を点で表し，その間の親子関係を線で結んだものを主個体 α の「木」T_α とする．たとえば数人の子を持つ個体は，それぞれの子の木に属していることからもわかるように，木同士も複雑に絡み合っていることが予想されるが，ここでは，まず一本の「木」の内部構造に着目する．

件のデータベースでは，古くは 18 世紀に誕生した個体を含めて 120,000 件以上の競走馬が登録されており，馬の名前は唯一に設定されている．(そのこと自体決してやさしいことではないと思われるが，登録しようとした馬名が重複しそうな場合には，名前に数字をつけることで区別している．) それぞれの名前を検索すると，{名前，誕生年，性別，父，母}というデータが得られる[5]．例えば，主個体 α として “Luke Skywalker” で検索すると，{“Luke Skywalker”, 1997, H[6] , “Skywalker”, “Border Country” } というリストが得られる．次に，(父) “Skywalker” および (母) “Border Country” で検索することによって “Luke Skywalker” の家系を順次遡ることができる．α から５世代遡るだけで，次のことが分かる（図２参照）．(i) α の父の母の母の父の父と，α の父の母の父の母の父は同一馬である．すなわち（α の父の母から見れば）３世代遡るだけで，木に重複して登場する個体がいる．(ii) α の父の母の父と，α の母の父の父の父の父は同一馬である．すなわち，ある個体が α の３世代前の先祖かつ５世代前の先祖ということであり，これは「世代」の重複（つまり仮定 A3 の破れ）を意味する．

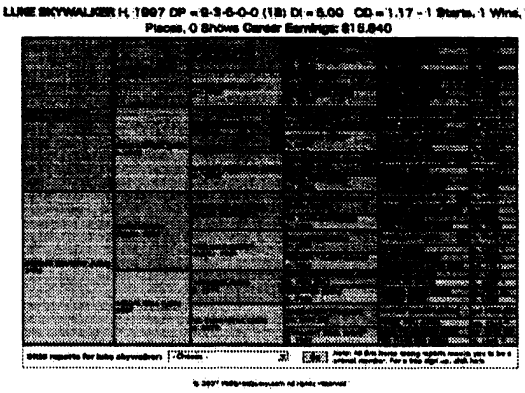

図 2. 競走馬のデータベースサイトの出力例．５世代前まで検索した結果．上欄は雄を，下欄は雌を表す．同じ濃淡の帯の個体は同一個体を示す．

以下，このデータベースから得られるデータとして，データ S（データベースに登録された馬の総個体数），データ R（個体集団の一般的な性質），データ T （主個体の木の性質）の３種類を考える．

---

[4] ある馬がサラブレッドであるためには，血統上の特定の条件を満足する必要があるが，その条件は年代とともに変化しており，一定ではない．ここでは，上記データベースに登録されている個体集団を取扱う．

[5] レースの成績や獲得賞金なども表示されるが，ここでは考慮しない．また死亡年は掲載されていない．子供のリストも得られるが，全ての子供がこのデータベースに登録される保証はないことに注意．

[6] “H” は雄を表す．“C” も雄（ただし，子馬）．雌は “M” あるいは “F” （子馬）である．

2.1 データS

データベースに登録された馬の総個体数：N_t : $1,200,000 < N_t < 1,300,000$. この値は，全てのデータの数を数えたわけではなく，ホームページに記載された記述「over 1.2 million」から見積もった．

2.2 データR

個体集団の一般的な性質：馬名からランダムに選んだ 4,349 頭に対し，{ 名前, 誕生年, 性別 } を調査した．選んだ個体に血縁関係があるかないかは無視している．このデータを用いて，個体集団の一般的な性質として誕生年分布を解析した．図3は，データRから読み取った誕生年分布であり，横軸は誕生年，縦軸は分布関数の対数をプロットしている．まず，雄より雌の方が個体数が多いことが分かる[7]．次に，それぞれ良く直線で近似できており，個体数は時間に対して指数関数的に増大していることを意味している．西暦 y 年に生まれた個体数 $N_Y(y)$ は，$N_Y(y)$ を用いて，

$$N_Y(y) = N_Y(2007) \cdot m_Y^{2007-y} \tag{1}$$

で良く近似できる．ここで m_Y は一年遡る毎にどれだけ個体数が少なくなるか[8] であり，雄，雌，合計いずれも約 0.980 である．

この数字と前小節で述べたデータベースに登録された総個体数 N_t は，データベースの開始年を y_0 とし，そこから指数 m_Y の指数成長が続いていると仮定すると，

$$N_t = \sum_{y=2007}^{y_0} N_Y(y) \tag{2}$$

で表される．このことから，$N(2007)$ を見積もることができる．N_t の範囲から $24,000 < N(2007) < 26,000$ となる．$y_0 = 1750$ でも，$y = -\infty$ でもこの値に大きな変動はない[9]．この値を式 (1) に代入すれば，その年に誕生した個体数を見積もることができる．

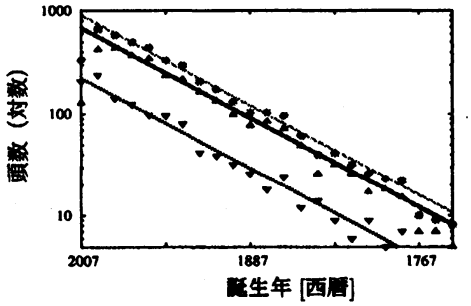

図3. ランダムサンプリングした 4349 頭の誕生年分布．●, ▼, ▲ はそれぞれ合計, 雄, 雌を表す．直線はそれぞれをフィットしたもの．

2.3 データT

特定の主個体の木の性質：図4は，ある主個体の家系図を 17 世代までさかのぼって描いたものである．縦軸は誕生年であり，横軸には「先祖座標」$b \equiv \dfrac{1}{\text{役割数}} \displaystyle\sum_{\text{役割}} \sum_{g=1}^{\text{世代}} \frac{(-1)^{\mathrm{sex}(g)}}{2^{g+1}}$ をとった．ここで g は主個体を 0 とし主個体から数えた世代を表す．$\mathrm{sex}(g)$ は g 世代個体の性関数であり，その個体が雄なら 0, 雌なら 1 と定義する．主個体の先祖座標を 0 とすると，その父母の世代は $g = 1$ であり，父の先祖座標は $(-1)^0/2^{1+1} = 1/4$, 母のそれは $(-1)^1/2^{1+1} = -1/4$ となる．祖父母は $g = 2$ であることから，父の父の先祖座標は $1/4 + (-1)^0/2^{2+1} = 3/8$ となる．同様に，父の母, 母の父, 母の母の先祖座標はそれぞれ $1/8, -1/8, -3/8$ で与えられる[10]．ある個体が木の中で複数の役割をはたしている（即ち主個体と複数の経路でつながっている）場合，その個体の先祖座標は，それぞれの役割に関して計算した先祖座標の平均を取るものとする．こうすれば，個体は唯一の座標を持ち，図の中で一点で表される．個体の生存期間は

[7] 正確には,「その年に生まれた個体の中でデータベースに登録された頭数は雄より雌の方が多い」である．これは個体の寿命が分からないからである．データベースには個体の死亡年が記載されていないので，個体の寿命がどのように分布しているか，またその性差の程度に関して推し測るのは難しい．しかし，次小節で述べる通り,「生殖年齢」に関して言えば，性差が少ないことを示唆するデータはある．

[8] すなわち一年あたりの増加率の逆数．

[9] そもそも $N(y)$ は自然数でなければならないので，あまり小さな y は無意味である．

[10] これらの値は，図1で各個体に用いられた横軸の値に一致する．

各記号から（少なくとも子の誕生年までは）上に延びていることになる．多くの個体を表すシンボルと親子関係を表現する線とが重なりあっていて見にくいが，傾向としては，母親系統（実線）は比較的独立した縦の線が見て取れるのに対し，父親系統（破線）は先祖座標で言う中程に密集している．これは父親の重複度合が大きいことを示唆している．以下，この木に関する解析を行う．

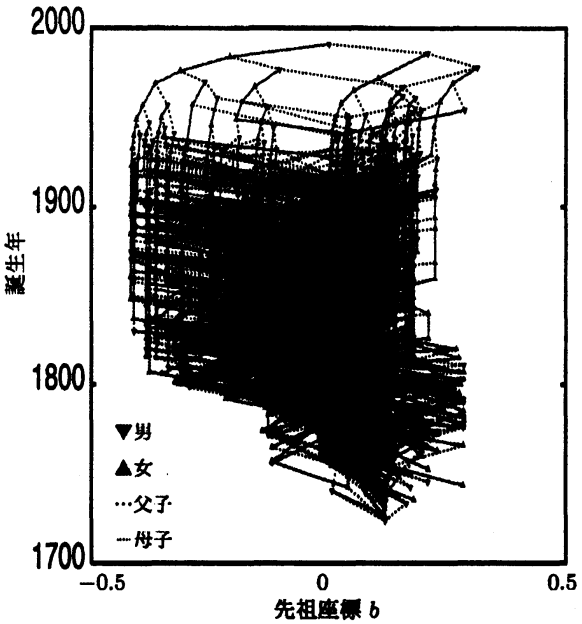

図 4. 一主個体から 17 世代まで遡った家系図．縦軸は誕生年，横軸は先祖座標を表す．

2.3.1 出産年齢と世代

親子関係に対して，子の誕生年と親の誕生年の差が，その親がその子を出産した年齢「出産年齢」である．図 4 の全ての親子関係に対して出産年齢を測定し，その分布を図 5 に示す．全体的な傾向は性別には寄らない．親子の平均値は，12 ± 1 年，標準偏差は 4.7 ± 0.2 年，最年少は 3 歳，最年長は 32 歳である．

誕生年分布から得られた各年代での個体数と木の中の個体数の比率を計算するために，出産年齢の平均値 12 年を 1 世代として採用しよう．すると，データ R で得られた個体数増加に関する諸データは，世代 $g = (y_0 - y)/12$ を用いて以下のように翻訳できる：世代 g に生まれた個体数

$$N_G(g) = N_G(0)\left(\frac{2}{m}\right)^g. \tag{3}$$

（ただし，主個体の誕生年 y_0 を，第 0 世代とし，過去に 12 年遡る毎に 1 世代増えるとした．）ここで分子の 2 は親の数を表し，分母の m は個体あたりの平均出産子数である．データ R での値を用いると，$m = 2.45$ となる．

2.3.2 重複の度合と先祖率

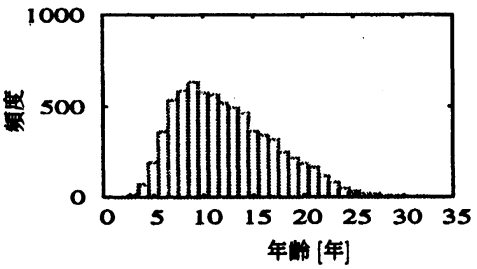

図 5. 図 4 の木に登場する個体の出産年齢分布

いよいよ，先祖の重複の程度を計算しよう．ある主個体の木を考え，世代 G にある正味の個体数を $A(G)$ とする．ただし，異なる役割を複数持つ個体の場合は，そのうちもっとも少ない値をその個体の世代とする．世代 G にある先祖の正味の数 $A(G)$ は，重複が無ければ $A(G) = 2^G$ だが，実際には重複があるのでそれよりも少ない．図 4 の木は，17 世代までのもので，のべ先祖数は $2^{17+1} - 1 = 262,143$ だが，正味の先祖数の総和は $\sum_{G=0}^{17} A(G) = 4,272$ に過ぎない．「のべ」に比べて極めて少数の先祖しかいないことが分かる．理論と比較するため，世代 G での先祖率 $A(G)/N(G)$ を図 6 に描いた．横軸は主個体の誕生年を 0 とした世代 G であり，縦軸は個体数の片対数グラフである．点線がのべ先祖数 2^G を表し，●印が図 4 の木から得られた正味の先祖数の実測値である．$G \sim 5$ あたりから，正味個体数（●）がのべ個体数（点線）から外れてはじめ，世代を遡るにつれその差が増大することがわかる．

2.3.3 理論との比較

図6には，理論との比較のため，集団全体の個体数 $N(G)$（破線）とDerridaの理論（実線）も描いてある．前者は2.1, 2.2節で見積もったものであり，傾き2.45の直線である．点線は（即ち，重複が無いとした場合の先祖の数）は，$G = 13$ 付近で集団全体の個体数 $N(G)$ を追い越しており，これが，イントロダクションで説明した先祖数のパラドクスである．実線は，Derridaらの理論を表しており $G \sim 20$ では，集団全体の個体数に漸近しているように見える．$m = 2.45$ の場合，漸近値は $A(G)/N(G) = 0.886$ となる．ところが，実測値では $G \sim 17$ ではDerridaらの理論曲線よりも一桁程小さい値となっており，同理論が妥当とは言いにくい．

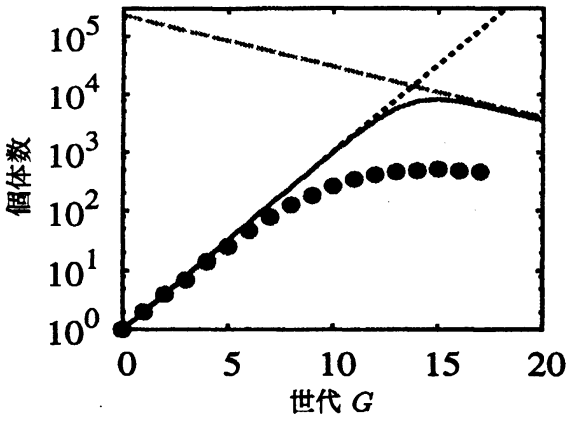

図6. 各世代での個体数と木の個体数．破線は集団全体の個体数 $N(G)$．点線はのべの先祖数 2^G．実線はDerridaらのモデルによる理論値．●は実測値 $A(G)$．

3 おわりに ～木から森へ

競走馬という実際の生物のデータを用いて17世代前までの家系図を構築した．ランダムに選別した個体の誕生年分布から個体数の指数成長を示し，総個体数データと組み合わせることによって，各年代の個体数を推定．家系図データと組み合わせることにより，主個体の先祖が各年代での個体数に占める割合—即ち先祖率を見積もった．これらの結果をDerridaの理論と比較し，相違点を見出した．まず，この原因に付いて考察してみよう．

(i) データの問題：たまたま特異なデータを選んだ可能性はないか，またデータは十分信頼できるか．前者に関しては，解析する木の数を増やすことである程度チェックできると思われるが，後者に関してはチェック方法（例えば，他のデータベースによる裏付け）がない現状では，信用するしかない．

(ii) 解析の問題：世代 G での個体数 $N(G)$ の見積り，総個体数 N_t の信頼性，データRの偏り，有限サイズ効果 $(N, G < \infty)$ など，誤差が入り込む要因は様々である．より多くのデータを解析することで誤差を減らすことは可能であろう．

(iii) 仮定の問題：Derridaらがそのモデルで採用した仮定のいくつかは明らかに破れている．世代は重複しているし，個体数や出産可能子数など性差も存在する[11]．両親の選択にも何らかの傾向（「淘汰圧」といっても良いかも知れない）があり，決してランダムではないと予想される．子数分布もポアソン分布からずれていることが予想される．

これらのことを考慮し，より多くのデータを注意深く解析すると同時に，新しい数理モデルを構築し解析することによって，家系図の重複の度合を正しく表現することが可能になると考えられる．

Derridaらは，家系図を特徴付ける量として，(非) 先祖率以外にも重複回数分布や血統のウェイト等といった量の分布や，MRCA(Most Recent Common Ancestor : 最近接共通先祖) のように，二主個体の木の絡み合いなどに関しても理論的な取り組みを行っている．競走馬の家系データは，これらの量についても測定可能である．これらのより詳細なデータに関しては稿をあらためて紹介したい．

家系図をたどるもう一つの方向として子孫方向が考えられる．図1（あるいは図4）は主個体にとって，先祖方向にたどったものであるが，系図は（時間がたてば）子孫方向にものびる．ただ，個体の子の数は親の数と違って一定ではない．子沢山の親もいれば，子がいない場合もあるので，枝の数（度数）も変化していると考えられる．これは後述するスケールフリー性と関連すると思われる．

集団全体に対する家系図—いわゆる「森」—の構造の解析は，よりチャレンジングな問題として残されている．いわゆる（方向性）ネットワークの構造という観点からとらえるためには，家系図全体の構造をシステマティックにとらえるデータが無いため，満足な観測結果を用いた議論は難しい．しかし，一本の

[11] ちなみに，母親の妊娠期間（すなわち出産可能間隔）は約1年であるとされている．

木に限定するならば，以下のようなことが考慮される．近親交配を避ける傾向からクラスタリング係数[12]は低いことが予想されるため，スモールワールドにはなりにくいと考えられる．

またスケールフリー性に関しては，雄と雌とで異なる可能性がある．まず，重要なパラメータとして可能出産数を考えよう．これは，一個体が出産することのできる最大の個体数であり，ネットワークの度数の最大値を制限する．論文の引用関係などは一種の家系図と考えることもできるが，論文は「死亡する」ことが無いので，無限に子孫を増やすことが可能である．しかし，(動物のように？) 寿命がある程度限定されている生物であれば，可能出産数には上限が存在する[13]．この上限は，例えば二つの時間スケールの比，出産可能期間／出産可能間隔で見積もることができるであろう．競走馬の場合，出産可能期間は雄雌いずれも 30 年程であるが，出産可能間隔は雌がおよそ 1 年に対して，雄のそれは非常に短い．つまり，雌はどんなに多くても 30 頭程度しか子を残せないが，雄の方はもっと多い (記録では 1,869 頭の親である雄がいる)．雄に関しては度数の上限はあるもののスケールフリーになっている可能性があると考えられる．祖先方向にたどる場合，一本の木では先祖の数は 2 と決まっているので，度数は一定である．しかし，子孫方向も考慮すると重複の結果，度数分布が偏っている可能性は否定できない．

系図を考える際，競走馬という生物集団は特殊ではないのか？という問いに関しては，「特殊であろう」と答えざるを得ない[14]．生殖行動は人間の管理下にあり，レースの成績や血統，経済状況による "淘汰圧" など，他の生物集団では考えにくい条件がいくつも考えられる．近親交配の度合も決して低くはないであろう．別の種類の生物集団と比較することができれば，競走馬がどの点でどの程度特殊であるかに答えることができると思われる．そのためにも，家系図構造の客観的な表現や特徴付けを行うことの意義は大きい．競走馬の平均出産年齢は 12 年と比較的長く，家系図も 20 世代以上遡ることはできていない．他の生物集団を用いて，より長い家系図を得ることができれば，漸近的な挙動にもより信頼性が増すであろう．

最後に，家系図に関する研究と他分野のかかわりについて触れたい．生物個体には遺伝子が乗っており，家系図というネットワーク上の遺伝子の動的挙動は，もともと遺伝学の対象である [5]．近年のネットワークに関する研究の発展が遺伝学に及ぼす影響も興味深い．また，家系図をもっと「遠く」から見たら何が見えるだろうか．そこには間違いなく，別の「種」との境界や分岐・融合，絶滅などが見えてくるであろう [6]．それは進化系統樹である．進化の歴史の巨視的な表現が進化系統樹がだとすれば，家系図はその微視的な表現であり，親子関係はその素過程なのではないだろうか．物理学において，巨視的な現象論である熱力学と微視的な統計力学との融合によって多くの成果が得られたように，生物学においても，物事をマクロな立場でとらえる生態学・人口学とミクロな立場で見る「家系図学」とが協力することで，生物の進化に対する新たな知見が得られればとも期待する [7]．

謝辞

長大な家系図が取得可能な対象として競走馬のことを指摘してくれた守田智氏に感謝します．また，原田崇広氏との有益な議論に感謝します．

参考文献

[1] "The Theory of Branching Processes", T. E. Harris, Springer-Verlag, (1963).

[2] S. Ohno, Proc. Nat. Acad. Sci. U.S.A., **93** (1996), 15276–15278.

[3] B. Derrida, S. C. Manrubia, & D. H. Zannette, Phys. Rev. Lett. **82** (1999) 1987–1990; B. Derrida, S. C. Manrubia, & D. H. Zannette, J. theor. Biol. **203** (2000) 303–315.

[4] D. L. T. Rohde, S. Olson, & J. T. Chang, Nature **431** (2004) 562–566.

[12] クラスタリング係数は，ある個体とつながっているもの同士が直接つながっている割合の高さであるが，親子関係では考えにくい．生き物のことだから例外はあるだろうけれども．

[13] クローン技術や万能細胞の可能性は考慮しない．また，植物の種子や胞子のように，明らかに異なるタイムスケールを持ち合わせている場合，家系図ネットワークの構造は異なる特徴を持つかも知れない．

[14] それでは「普通の」生物とはなんだろうか？ご存知の方がいれば是非教えて頂きたい．

[5] "Biology and Language : an introduction to the methodology of the biological science including medicine.", J. H. Woodger, Cambridge University Press, Cambridge (1952).

[6] R. J. O'Hara, American Zoologist **34** (1994), 12–22.

[7] 「生物系統学」, 三中伸宏, 東京大学出版会, (1997).