令和 4 年度　修士論文

ループ構造に着目した家系図の構造解析

Structural analysis of family trees focusing on a loop structure

大阪府立大学大学院理学系研究科物理科学専攻

非線形物理研究室

学籍番号 2210302001

赤石　大夢

令和 5 年 3 月 3 日

概要

本研究の目的は、有性生殖を行う生物の家系図が本質的に有する複雑な構造を定量化し、比較することである。そのため、家系図のネットワークを特徴付ける構造として祖先ループを定義し、その統計的な性質に着目した。簡単な数理モデルを用いて仮想的な家系図を作製し、累積祖先ループ数を測定した。また、特定の条件下で累積祖先ループ数のパラメータ依存性を理論的に導き出すことに成功した。解析結果から累積祖先ループ数は探索世代に対して指数関数的に増大し、初期個体数に応じて減少することが、理論および数値計算から示された。次に、実際の家系図として競走馬の家系図に着目し、同じ年に生まれた個体群の累積祖先ループ数を測定した。その結果、モデルと同様に探索世代に対して指数関数的に増大するが、累積祖先ループ数はモデルと比べ多いことが判明した。さらに雄雌対称性を持たないようにモデルと理論を拡張した。累積祖先ループ数を測定した結果、理論と数値計算が無矛盾であることがわかった。

目次

1章　序論

ネットワークとは、頂点と呼ばれるモノとそれらをつなぐ辺から構成されており、モノとモノ同士のつながりを数学的に表現するときに用いられる。実在するネットワークの例として人間同士の交友関係がある。図 1.1 は交友関係を表したネットワークの例である。この場合頂点が個人で、辺が交友関係に対応する。ネットワークには次数と距離がある。次数は頂点から出る辺の数であり、距離は頂点同士を結ぶ経路の中で最も辺の数が少ないものの辺の数である。図 1.1 の頂点 A の次数は、A から出る辺の数であることから 3 である。また A と F の距離は、A と F を結ぶ経路の中で最も辺の数が少ない経路（A、E、F を結ぶ経路）の辺の数であるから 2 である。

実在のネットワークは他にも様々なものがある。例えば路線図や俳優の競演関係がある。前者は頂点が駅で辺が線路に対応し、後者は頂点が個人で辺が共演関係に対応している。

ネットワークに関する研究は多く行われている。Watts と Strogatz は規則的なネットワークとランダムなネットワークの間を連続的につなぐ方法を提案し、両者の間に新しいクラスのネットワークがあることを示した[1]。彼らはそのネットワークをスモールワールドネットワークと呼んだ。それは規則的なネットワークの各頂点から出る各辺を確率 p（0≦ p ≦ 1）で張り替えることでできたグラフである。 p ＝0 の時は規則的なグラフであり、また p＝1 の時はランダムなグラフである。彼らはネットワークの二頂点間の平均距離とクラスター係数に注目した。二頂点間の平均距離を L(p)、クラスター係数を C(p)、頂点数と次数をそれぞれ n と k とした。p が 0 に近づく、つまりネットワークが規則的になると L は n/2k に近づき、C は 3/4 に近づくことを示した。一方、p が 1 に近づく、つまりネットワークがランダムになると L は ln(n)/ln(k)、C は k/ n になることも示した。これらのことから、規則的なネットワークは高度にクラスター化されており、頂点数 n が増えるにしたがってネットワークの平均距離 L は線形的に増加すると主張した。一方、ランダムなネットワークはクラスター化が少なく、頂点数 n が増えても平均距離 L は対数的にしか増加しないことを示した。また L(p)/L(0)と C(p)/C(0)の変化を見ると、p の値が小さい場合、クラスター

係数の比の値はあまり変化せず、pが1に近づくにつれ小さくなった。一方、平均距離の比は p の値が小さい場合、急激に小さくなることがわかった。これらからスモールワールドネットワークはクラスター性が高く、かつネットワークの平均距離が短いと主張した。

ネットワークについての様々な研究の中で、本研究は特に家系図ネットワークに着目した。家系図ネットワークとは頂点を個体、辺を親子関係で構成したものである。家系図ネットワークには一般的なネットワークと異なり、過去と未来が存在し、次数は過去方向に必ず2であるという特徴がある。また、親子関係を結ぶ家系図ネットワークは、家系図を過去方向にのみ辿る、あるいは未来方向にのみ辿るといったように、辿る方向を一方向に絞ることで有向非巡回グラフ（directed acyclic graph 以下、DAG）の一種とみなすことが出来る。DAG の一例としては、学術論文の参照ネットワークが挙げられる。参照ネットワークでは、論文の価値は一様ではなく、参照ネットワークの頂点（論文）の中には、後の論文に大きな影響を与える論文が存在すると考えることができる。Gualdi らは、ある論文からランダムウォークでリファレンスを辿り、別のある論文へ到達する確率を用いて、ネットワーク上の他の論文に大きな影響を与えている論文を探す方法を提案した[2]。図1.2は論文の”子孫”数と影響度を表したものである。この子孫数とはある論文を参照した論文の数に対応している。子孫数が大きくなると影響度も大きくなることがわかり、特に影響度の大きい論文はグレーの範囲より上にプロットされていることがわかる。

家系図に関する先行研究としては、まず Derrida らによるものが挙げられる[3]。これは、中立なモデルを用いて、ある着目した個体に対して、その先祖らそれぞれが与える寄与の大きさを求めたものである。用いられたモデルはDerrida モデルと呼ばれ、本研究で用いるモデルもこの先行研究のモデルを参考にしている。他に Ikuta は親子関係を未来方向にたどることを考え、ある着目した個体に対して、その子孫らにあたえる寄与を 2 種類定義し、それらについて解析している[4]。

ここで、家系図ネットワークが本質的に持つ構造の複雑さについて簡単に説明しておこう。例えば図1.3はある個体の三世代前までの全祖先を表示したものである。このように有性生殖をおこなう生物における主個体の祖先数は、親が 2、祖父母が 4…と 2 の べき乗で増加していくことがわかる。つまり主個体の G 世代前の祖先数をN_Gとすると、

$$N_G = 2^G \tag{1.1}$$

となる。しかし、有性生殖する家系図には同じ個体が複数回登場することがあり、異なる祖先数は2^Gより少なくなる場合がある。もし、図 1.3 において祖先γ_1とγ_2が同一個体γであったとすると、三世代前の異なる祖先数は 7 となり$2^3 = 8$よりも少なくなる。実際の家系図では、この個体γのようにαの父の母の父かつαの母の父の父といった複数の役割をもつ祖先個体が数多く存在していると考えられる。

このような場合、家系図ネットワークにおいては、複数回登場する個体への経路がループ構造、すなわち同じ頂点を通らずに元に戻ってくる経路を形成する場合がある。図 1.4 は図 1.3 の家系図でγ_1とγ_2が同一個体γだった場合の家系図であるが、緑線のようにαとγを通るループ構造があることがわかる。このように家系図のネットワークにはループ構造が多く存在し、複雑な構造が形成されていると考えられる。

本研究の目的はこの複雑な構造を定量的に捉えることである。特にループ構造の中でも祖先ループと呼ばれるものを定義し、その数の分布によって家系図の構造を特徴づける。まずモデルによって仮想的な家系図を構築し、それを用いて解析方法を紹介する。解析は、祖先ループの統計的性質に関する理論式の導出を行ったあと、家系図モデルから統計量を実測し、両者を比較する。実際の家系図や拡張モデルにも解析方法を適用し比較する。その結果判明した仮想的な家系図と実際の家系図の違いについて述べ、その原因に関する考察を行う。

本論文の構成は以下の通りである。第 2 章で Derrida らの先行研究と Ikuta の先行研究を紹介した後 Derrida モデルの具体的な実装方法を説明する。第 3 章では祖先ループを定義し、それを用いて Derrida モデルを対象として行った解析結果を報告する。第 4 章では競走馬のデータを用いた解析結果を報告する。第 5 章では Derrida モデルを拡張した家系図モデルに対する解析結果を報告する。第 6 章で本研究のまとめと今後の課題について述べる。

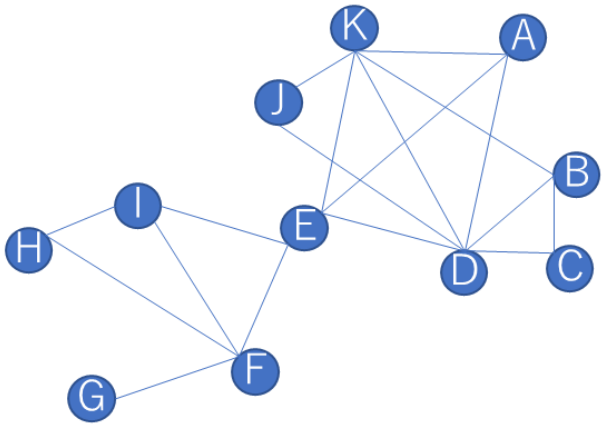

図 1.1　交友関係を表したネットワークの例。A~K は名前。頂点が個人、辺が交友関係。

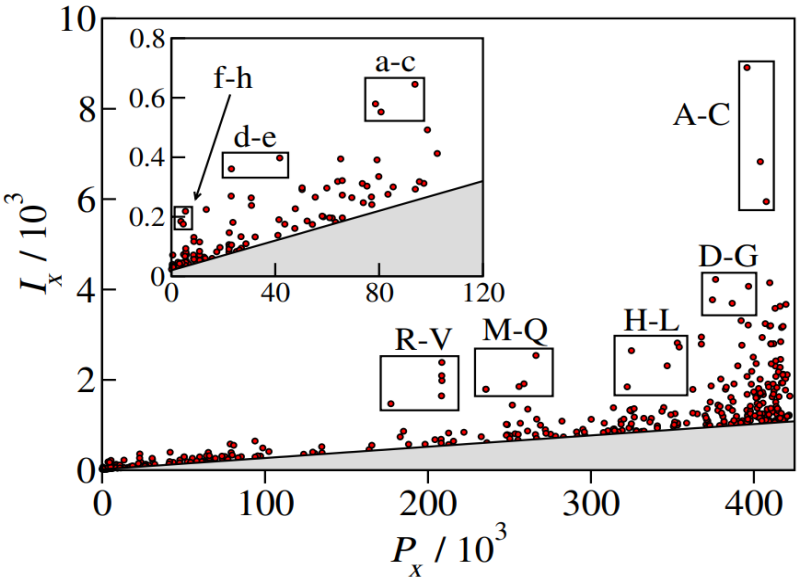

図 1.2　論文の”子孫”数と影響度。Gualdi, et al.2011 より転載。

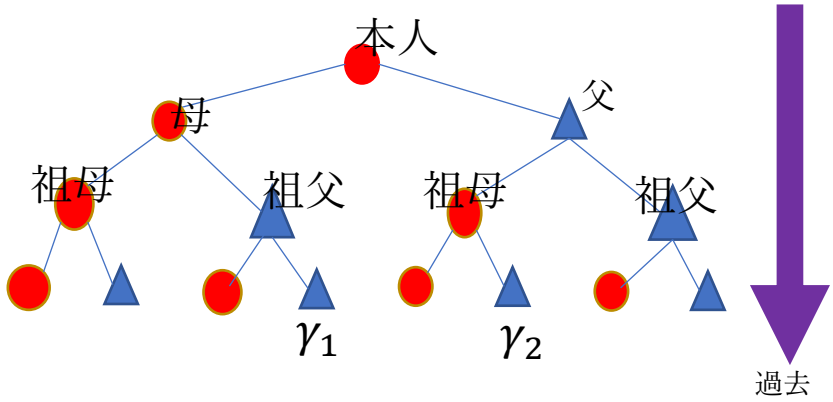

図 1.3　ある個体の家系図ネットワークの例。三世代前までの全祖先を表示。γ_1とγ_2が同一個体の場合、図 1.4 のようになる。

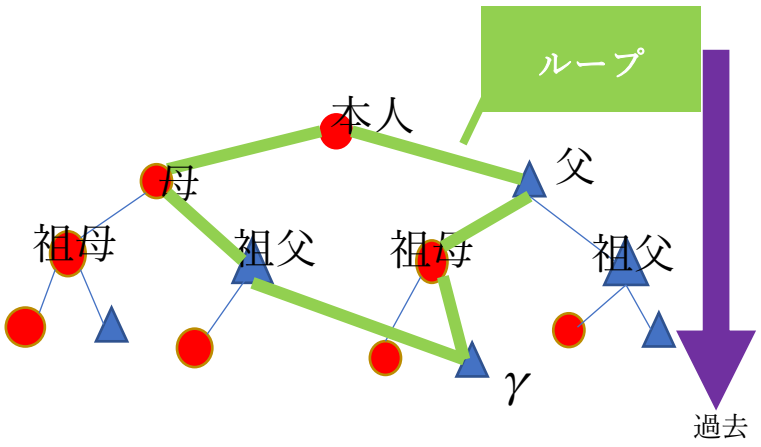

図 1.4　図 1.3 の家系図でγ_1とγ_2が同一個体γであった場合の家系図。緑線はループ構造。

2 章　Derrida モデル

本研究において、構造解析の主な対象としたのは、Derrida によって提案された仮想的な生物集団によるモデル家系図である。本章では、この Derrida モデルに関する先行研究として Derrida ら自身によるもの[3]と、Ikuta によるもの[4]を紹介した後、その詳細を説明する。

2.1 先行研究

Derrida らは以下の 4 つの特徴を持つ中立的な仮想的生物集団のモデルを提案した[3]。(i)時間は離散化された世代で表され、各個体の寿命は一世代である。(ii)個体には性別があり、同世代の中からランダムに選ばれた雄雌のペアから子が次の世代に生まれる。(iii) 子供の数は Poisson 分布に従う。(iv) 個体数および子数分布は雄雌同じである。以下これを Derrida モデルと呼ぶ。Derrida モデルの詳細な説明および実装方法については、次節で述べる。Derrida らは、家系図に現れる個体がその祖先個体から受け継ぐ遺伝子の量であるウェイトに着目し、提案したモデルにおいてその分布を理論的および数値的に解析した。

ウェイトとは血縁関係の近さを表現するとも言える。注目する子孫個体 α を主個体と呼び、主個体 α の先祖個体 γ のウェイトは以下の式によって定義される。

$$\mathrm{w}(\alpha,\alpha) \equiv 1 \tag{2.1}$$

$$w(\gamma,\alpha) \equiv \sum_{\gamma\text{の子}\gamma'} \frac{w(\gamma,\gamma')}{2} \tag{2.2}$$

α 自身のウェイトを 1 とし、その祖先に対して、親のウェイトは、子のウェイトの総和の半分となる。祖先でなければ、ウェイトの値は 0 である。例えば図 2.1 では個体 γ の主個体 α に対するウェイトは $\mathrm{w}(\gamma,\alpha) = 2/4$ となる。

Derrida らは主個体 α から十分離れた世代に属する γ らが持つウェイトの分布が、ウェイトが微小な範囲においてべき則に従うことを示した。またべき指数の平均子数依存性も見積もっている。(図 2.2)

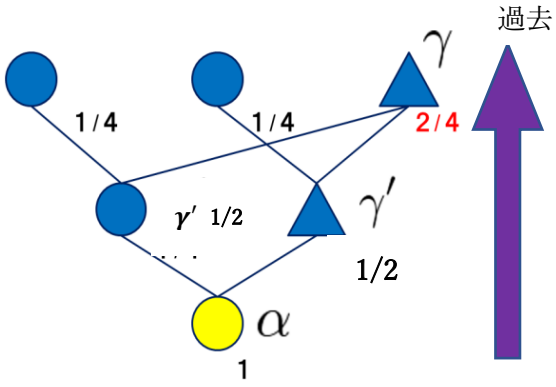

図 2.1　ウェイトの計算例。個体の右下の値はそれぞれの持つウェイトを表す。先祖 個体 γ の主個体 α に対するウェイトは $w(\gamma, \alpha) = 2/4$ となる。Ikuta 2014 より改変。

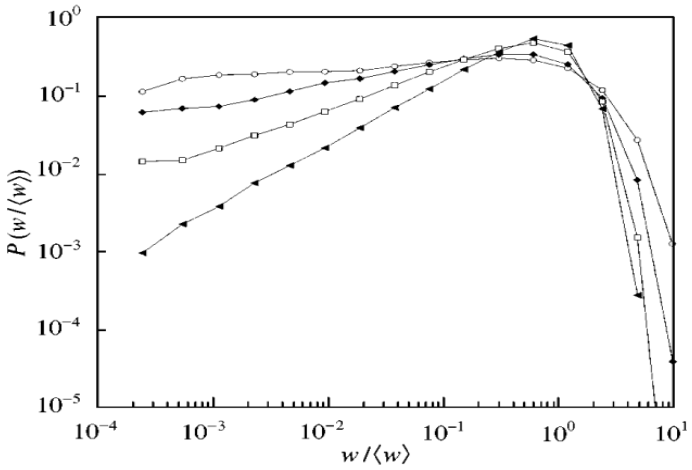

図 2.2　Derrida モデルのウェイトの分布。Derrida, et al. 2000。

家系図を過去方向にたどった Derrida に対して Ikuta は親子関係を未来方向にたどることを考え、ウェイトに対応するような量を二種類定義した[4]。すなわち血縁関係の近さを表現する量である関係性rと子孫個体 α への到達確率を表現する量である遺産 h である。それらの量を用いて、Derrida モデルで構築した家系図を解析した。主個体 α の先祖個体 γ の関係性rは以下の式によって定義される。

$$r(\gamma,\gamma) \equiv 1 \tag{2.3}$$

$$r(\gamma,\alpha) \equiv \sum_{\alpha\text{の両親個体}\alpha'} \frac{r(\gamma,\alpha')}{2} \tag{2.4}$$

すなわち着目個体を 1 とし、子に親の持つ量の和の半分が与えられる。例えば図 2.3 の主個体 α の祖先個体 γ に対する関係性の値は $r(\gamma,\alpha) = 2/4$ となる。図 2.4 に表されるように関係性とウェイトの分布はよく一致しており、またどちらも値が微小な範囲においてべき則を示している。

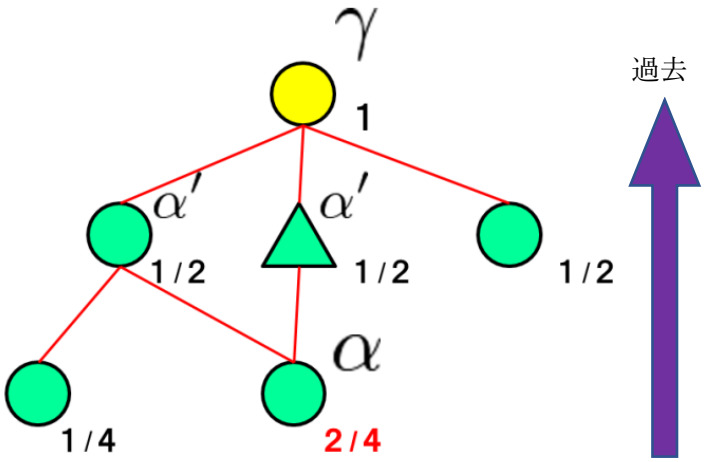

図 2.3　関係性の計算例。個体の右下に書かれた数字はその個体の持つ関係性の値を表す。Ikuta 2014 より改変。

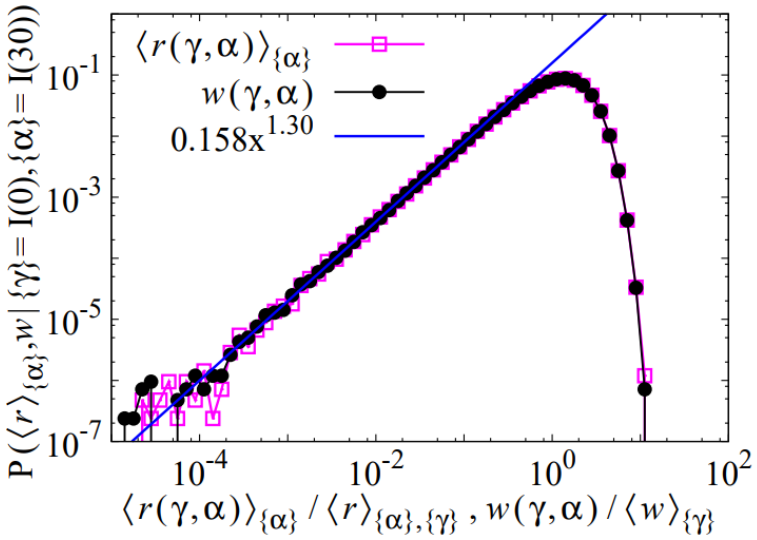

図 2.4　関係性の平均値の分布(紫四角)。黒丸はウェイトの分布。直線は $\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ の分布の 10^{-4}~10^{-1} の範囲でフィッテングをしたものである。Ikuta 2014 より転載。

これに対して、着目個体 γ の子孫個体 α の持つ遺産 h は以下の式によって定義される。

$$h(\gamma,\gamma) \equiv 1 \tag{2.5}$$

$$h(\gamma,\alpha) \equiv \sum_{\alpha\text{の両親個体}\alpha'} \frac{h(\gamma,\alpha')}{N_{\alpha'}} \tag{2.6}$$

着目する祖先の遺産を 1 とし、個体α'の子αには、α'の遺産$h(\gamma,\alpha')$をα'の子の数$N_{\alpha'}$で等分した量を与える。役割が重複した子孫にはすべての直接の両親から与えられる和になる。計算例を図 2.5 に示す。

結果は図 2.6 である。それぞれの個体が十分離れた世代に与える遺産の分布は微小領域においてべき則を示すことが分かる。べきの指数は関係性の指数と異なっている。べきの指数に関する理論的な考察はなされていない。

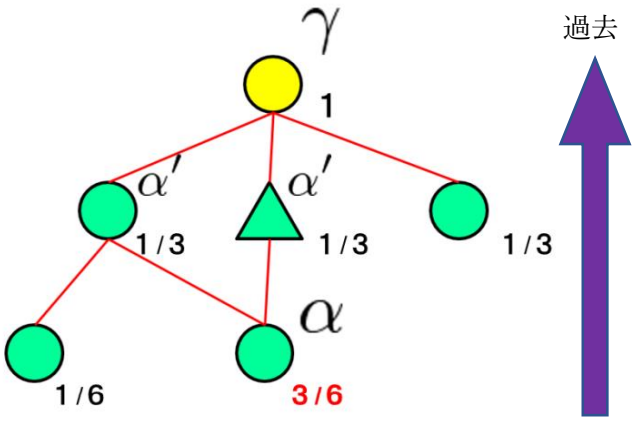

図 2.5　遺産の計算例。個体の右下に書かれた数字はその個体の持つ遺産の値を表す。Ikuta 2014 より改変。

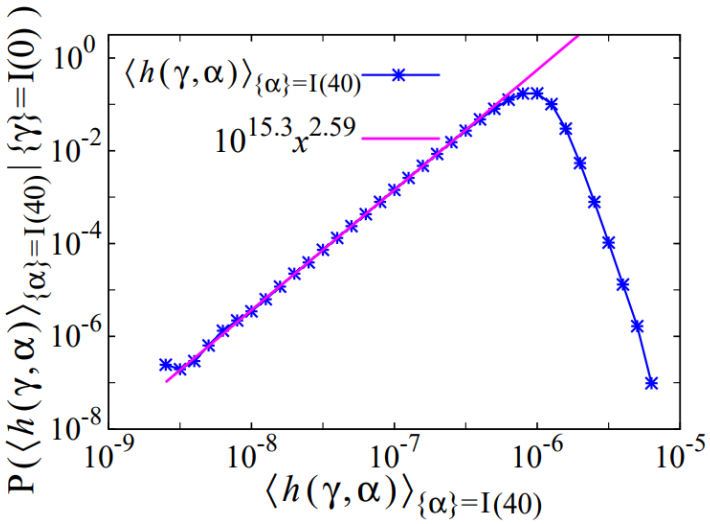

図 2.6　遺産の平均値の分布。直線は得られた分布をその微小領域でフィッティングしたものである。Ikuta 2014 より転載。

2.2 Derrida モデル

本節では Derrida モデルの詳細と実装方法を説明する。
Derrida モデルは以下の条件を満たす。

A.1 初期個体数は N 体（雄 N/2 体、雌 N/2 体）とする

A.2 世代は最も過去から G=0,1,2…とする。各個体は 1 世代しか生きられない

A.3 子は必ず親の次の世代で生まれる

A.4 配偶制度は一夫一妻制とする

A.5 配偶する相手は同じ世代の個体からランダムに 1 体選ぶ

A.5.1 ペアを組む数は個体数が少ない方の性別の数に合わせる

A.5. 2 ペアを組めなかった個体は子供を産まないものとする

A.6 子の性別は雌 50%、雄 50%の確率で振り分ける

A.7 各世代の個体数が大きい場合、子数 c の確率密度分布p_cは雄雌いずれもパラメーターλの Poisson 分布

$$p_c = \frac{\lambda^c \, e^{-\lambda}}{c!}$$

に従う。

A.1 かつ A.7 のようにどの世代でも雄と雌の個体数が 1：1 かつ子数分布も等しい場合、家系図は雄雌対称であるという。A.2 かつ A.3 の条件がある家系図を世代同期型と呼ぶ。A.7 のλは個体数の増減に関しており、λ＝2 のとき個体数はおよそ一定となる。個体数がおよそ一定となるモデルを特に単純 Derrida モデルと呼ぶ。本節ではこの単純 Derrida モデルを取り扱う。

これらの条件をもとに以下のプロセスで仮想的な家系図を作製した。

B.1 G＝0 世代の個体を雌 N/2 体、雄 N/2 体用意する

B.2 配偶制度の条件 A.5、A.5.1、A.5.2 をもとにペアを組む

B.3 ペアごとに Poisson 分布に従った乱数を振り、子数を決定し
　　性別を振り分ける

B.4 Gmax 世代まで同様に繰り返す

上記の条件とプロセスをもとに作成した家系図の例が図 2.7 である。この図では世代ごとの人口は同じであるが、厳密には人口の変動が見られる。というのは、上記の条件 A.5 と A.7 に基づいて家系図を作成すると、各世代で子数が変わり、またペア数も変わるからである。本研究では λ を 2 より少し大きな値をとることで、人口の変動をおさえた。

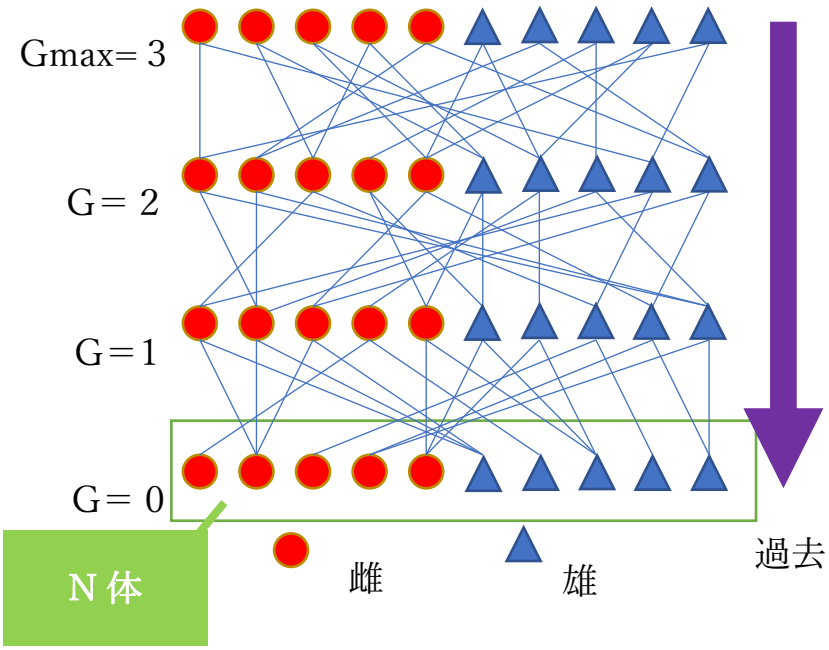

図 2.7　Derrida モデルによる家系図の例。

3 章　祖先ループに着目した解析

本章では、家系図ネットワークの構造解析において着目した祖先ループを導入した後、その分布に関する理論式を示す。得られた結果を Derrida モデルに条件を加えたものに適用し、理論式の妥当性を示す。

3.1 祖先ループ

本研究では、有向非巡回グラフに特有な構造に着目した。以下、この構造を祖先ループと呼ぶ。祖先ループは、ある着目個体から両親の共通祖先までさかのぼる経路のペアから構成されるループと定義される。有性生物の家系図ネットワークでは任意の個体に対してその家系図を十分にさかのぼることで、祖先ループが存在すると考えられる。これは以下のような考察に基づく。有性生物の各個体の祖先の数は、一世代前は母と父の 2 体、二世代前は 4 体、三世代前は 8 体、と 2 の世代乗で指数的に増えていく。これを外挿すると祖先の数は必ずある年の総個体数を上回る事になってしまう。これは全ての先祖を別々の個体とし、重複した個体を考慮していないことが原因である。つまり必ず重複した個体が存在する。重複個体の中には両親の共通祖先がおり、その共通祖先までさかのぼる経路のペアがループをなせば祖先ループが存在することになる。

祖先ループを実際に見てみよう。図 3.1 は個体 α の祖先のネットワークの一部である。γ_1、γ_2は α 個体の両親の共通祖先である。図の水色と緑色の経路のペアは祖先ループである。このペアは α 個体から両親の共通祖先であるγ_1までさかのぼる経路でループを構成しているからである。次に図の水色と橙色の経路のペアは祖先ループでない。個体 z から未来方向にたどる経路があり、これは祖先ループの定義から外れているからである。また、黄色と紫色の経路のペアも祖先ループでない。このペアは α 個体から両親の共通祖先であるγ_2までさかのぼる経路であるが、γ_1を 2 回通っているので、ループの定義から外れている。

祖先ループは、後述するように遺伝学における近郊係数と密接に関係している。祖先ループの数を調べることで家系図ネットワークの構造を定量的に表すことができるのではないかと考えた。

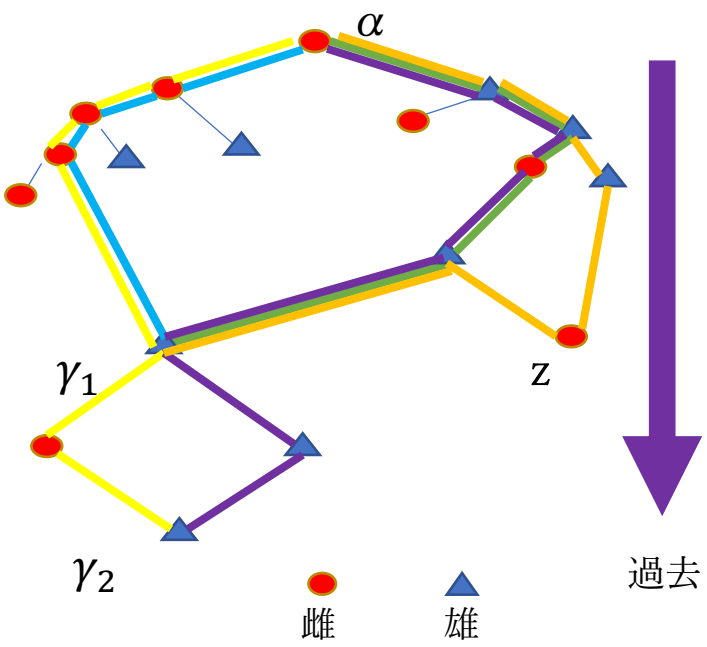

図 3.1　祖先ループの例。γ_1、γ_2はαの両親の共通祖先

3.2　累積祖先ループ数に関する理論式

本節では、祖先ループ数の統計的な性質として、累積祖先ループ数の世代依存性に関する理論式を導出する。まず Derrida モデルによる家系図の集団を考えよう。すなわち、世代同期型で世代は G = 0, 1, 2,…, Gmax,とする。性別 σ = ♀, ♂ に対して決まる個体数$N^{\sigma}(G)$とする。子数 c の確率密度関数をp_c^{σ}とすると子数分布は$p_c^{\sigma}N^{\sigma}(G)$であり、親子関係はランダムに決められている。本節では、ある一つの家系図に対して，G=Gmax の個体の中からランダムに選ばれた個体 α に着目する。α から先祖方向に遡った世代数を探索世代 g と呼び、g 世代前まで遡った時の累積祖先ループ数 K(g)の平均値を理論的に求める。探索世代 g と世代 G は g=Gmax－G の関係がある。したがって、 g は G と異なり、最も未来の世代を 0 とし、過去方向に増加することに注意する必要がある。

この理論では二種類の平均が用いられる。個体数$N^{\sigma}(G)$と子数確率密度分布$p_c^{\sigma}N^{\sigma}(G)$が同じ（だが、乱数の種が異なる）家系図の集合、およびその中の α 個体を母集団とした場合の量Qの平均を$\bar{Q}$で表す。これに対して、子数確率密度分布$p_c^{\sigma}(g)$に対する量 Qの平均を $\langle Q\rangle^{\sigma}$で表す。すなわち

$$\langle Q \rangle^{\sigma} = \sum_{c=0}^{\infty} Q p_c^{\sigma}(g) \tag{3.1}$$

である。

本節の目的は累積祖先ループ数の平均 $\bar{K}(g)$を解析的に求めることである。探索世代 g=0 の個体αからさかのぼり、先祖方向にちょうど g 世代前の祖先で閉じる祖先ループ数の平均を$\bar{k}(g)$とすると、累積祖先ループ数の平均は

$$\bar{K}(g) = \sum_{g'=0}^{g} \bar{k}(g') \tag{3.2}$$

で与えられる。祖先ループができるのは g=2 以上なので、和をとる下限は 2 にしても良い。以下、$\bar{k}(g)$を求める。ちょうど g 世代で閉じる祖先ループを構成する可能性がある経路のペアは、αの母から g -1 世代さかのぼる経路2^{g-1}本と α の父から g-1 世代までさかのぼる経路2^{g-1}本のペアなので、合計 $2^{g-1} \times 2^{g-1} = 4^{g-1}$ だけある。これらの経路のペア全てについて g 世代で初めて閉じる確率を求める。αの母から g-1 世代さかのぼる経路の一本を ξ とし、α の父から g -1 世代さかのぼる経路の一本を η とする。後述するように ξ と η は経路上の祖先の性別の組で指定する。ξ と η のペアが g 世代で初めて閉じる確率を$\kappa_{\xi\eta}(G)$とする。この確率は、個体数$N^{\sigma}(g)$ と子数分布 $p_c^{\sigma}(g)$は同じ（だが、乱数の種が異なる）家系図の集合およびその中の α 個体を母集団とした場合の確率であることに注意する。各ペアについて祖先ループをなす確率が$\kappa_{\xi\eta}(g)$なので、

$$\bar{k}(g) = \sum_{\xi,\eta} \kappa_{\xi\eta}(g) \tag{3.3}$$

となる。和は4^{g-1}通りある。$\kappa_{\xi\eta}(g)$を計算する方針は以下の通りである。ξとηは $g' = 1, 2, \cdots, g-1$ までは閉じずに、$g' = g$で初めて閉じるので、ξとηが世代 g で同じ祖先個体を通る確率を$q_{\xi\eta}(g)$とすると

$$\kappa_{\xi\eta}(g) = \Big(1 - q_{\xi\eta}(2)\Big)\Big(1 - q_{\xi\eta}(3)\Big)\cdots\Big(1 - q_{\xi\eta}(g-1)\Big) q_{\xi\eta}(g) \tag{3.4}$$

$$= \prod_{g'=2}^{g-1} \left(1 - q_{\xi\eta}(g')\right) q_{\xi\eta}(g) \tag{3.5}$$

となる。式の右辺に$\left(1 - q_{\xi\eta}(1)\right)$がないが、これは$q_{\xi\eta}(1) = 0$すなわち 1 世代だけ遡っても、同一個体になりえないことから来ている。(1 世代前の祖先は母と父である。)

ξ 上の祖先個体を $\xi(g')$, $g' = 1, 2, \ldots, g$ とし、η 上の祖先個体を $\eta(g')$, $g' = 1, 2, \ldots, g$ とする。$\xi(1)$は α の母であり、$\eta(1)$は α の父である。$\xi(2)$は α の母方の祖父母のいずれかであり、$\eta(3)$は α の父方の曾祖父母 4 名のいずれかである。$q_{\xi\eta}(g)$を計算する上で、祖先個体 $\xi(g')$の性と祖先個体 $\eta(g')$の性が二つの点で重要である。まず $\sigma(\alpha)$で個体 α の性別を表すことにする。$\sigma(\xi(g')) \neq \sigma(\eta(g'))$ならば、世代 g で経路のペアが閉じることはない。言い換えると経路のペアが閉じるためには $\sigma(\xi(g')) = \sigma(\eta(g'))$ が必要条件となる。実際$g' = 1$ を考えると $\xi(1)$は母親で, $\eta(1)$ は父親なので、性別は異なり、そこでは閉じない。さらに、雄雌対称性が破れている場合は、$\sigma(\xi(g')) = \sigma(\eta(g')) = ♀$の場合と $\sigma(\xi(g')) = \sigma(\eta(g')) = ♂$の場合で、閉じる確率が変わる。これらのことに注意して$q_{\xi\eta}(g)$を求める。

子数確率密度分布$\{p_c^\sigma(g')\}$の元でのランダムな親子関係とは, $g' - 1$ 世代の個体の母親は g' 世代の雌から$g' - 1$世代に向かって伸びている全ての辺$N^♀(g')\sum_c cp_c^♀(g') = N^♀(g')\langle c\rangle^♀$の中からランダム に選び、$g' - 1$ 世代の個体の父親は g'世代の雄から $g' - 1$ 世代に向かって伸びている全ての辺$N^♂(g')\sum_c cp_c^♂(g') = N^♂(g')\langle c\rangle^♂$の中からランダムに選ぶということを意味する。今 ξ と η は $0 \leq g' \leq g - 1$ までは閉じない状況を考えているので $\xi(g' - 1)$と $\eta(g' - 1)$ は異なる個体である。したがって $\xi(g')$と $\eta(g')$が性別σの同一個体になる確率$q_{\xi\eta}^\sigma(g')$は

$$q_{\xi\eta}^\sigma(g') = \frac{N^\sigma(g')\sum_c {}_cC_2 p_c^\sigma(g')}{{}_{N^\sigma(g')\langle c\rangle^\sigma}C_2} \delta_{\sigma\sigma\left(\xi(g')\right)} \delta_{\sigma\sigma\left(\eta(g')\right)} \tag{3.6}$$

$$= \frac{N^\sigma(g')\langle c(c-1)\rangle^\sigma}{N^\sigma(g')\langle c\rangle^\sigma(N^\sigma(g')\langle c\rangle^\sigma - 1)} \delta_{\sigma\sigma\left(\xi(g')\right)} \delta_{\sigma\sigma\left(\eta(g')\right)} \tag{3.7}$$

$$= \frac{\langle c^2\rangle^\sigma - \langle c\rangle^\sigma}{\langle c\rangle^\sigma(N^\sigma(g')\langle c\rangle^\sigma - 1)} \delta_{\sigma\sigma\left(\xi(g')\right)} \delta_{\sigma\sigma\left(\eta(g')\right)} \tag{3.8}$$

となる。式(3.6)の分子は世代g'で性別σの一個体から伸びている親子関係の中から 2 本選択する場合の数であり、分母は世代g'で性別 σ の全ての個体から伸びている親子関係（合計$N^{\sigma}(g')\langle c\rangle^{\sigma}$本）の 中から 2 本選択する場合の数である。最後の係数はいずれも Kronecker のデルタであり、$\xi(g')$と $\eta(g')$の性別が σ に一致した時のみ 1 となる。性別ごとに分けて書くと、

$$q_{\xi\eta}(g')=\begin{cases}\dfrac{\langle c^2\rangle^{♀}-\langle c\rangle^{♀}}{\langle c\rangle^{♀}(N^{♀}(g')\langle c\rangle^{♀}-1)} & if\ \sigma\big(\xi(g')\big)=\sigma\big(\eta(g')\big)=♀\\ \dfrac{\langle c^2\rangle^{♂}-\langle c\rangle^{♂}}{\langle c\rangle^{♂}(N^{♂}(g')\langle c\rangle^{♂}-1)} & if\ \sigma\big(\xi(g')\big)=\sigma\big(\eta(g')\big)=♂\\ 0 & otherwise\end{cases} \tag{3.9}$$

となる。

もっとも単純な例として、λ＝2 の単純 Derrida モデルに適用してみよう。このとき個体数と子数確率密度分布はそれぞれ$N^{\sigma}(g')=N,\{p_c^{\sigma}(g')\}=\{p_c\}$のように σ, g' によらない定数となる。$\langle c\rangle^{\sigma}=\langle c\rangle=2,\langle c^2\rangle^{\sigma}=\langle c^2\rangle$とおくと

$$q_{\xi\eta}(g')=\begin{cases}\dfrac{\langle c^2\rangle-2}{2(2N-1)}\equiv q & if\ \sigma\big(\xi(g')\big)=\sigma\big(\eta(g')\big)\\ 0 & otherwise\end{cases} \tag{3.10}$$

となり q は N には依存するが g に依存しない。この時、式(3.5)は

$$\kappa_{\xi\eta}(g)=\prod_{g'=2}^{g-1}\Big(1-q\delta_{\sigma(\xi(g'))\sigma(\eta(g'))}\Big)\times q\delta_{\sigma(\xi(g'))\sigma(\eta(g'))} \tag{3.11}$$

となる。4^{g-1}通りの経路のペアに対して、$\sigma\big(\xi(g')\big)=\sigma\big(\eta(g')\big)$、すなわち$\xi(g')$の性別と$\eta(g')$の性別が一致する確率は 1/2 なので、(3.11)式は

$$\kappa_{\xi\eta}(g)=\Big(1-\frac{q}{2}\Big)^{g-1}\times\frac{q}{2} \tag{3.12}$$

となる、これは停止確率 q/2 の幾何分布である。(3.3) 式に代入すると

$$\bar{k}(g) = 4^{g-1} \times \left(1 - \frac{q}{2}\right)^{g-1} \times \frac{q}{2} \tag{3.13}$$

となる。N が大きい場合、$p_c = \frac{2^c e^{-2}}{c!}$と近似できる。$\langle c^2 \rangle = 6, \langle c \rangle = 2$なので，$q = \frac{\langle c^2 \rangle - \langle c \rangle}{\langle c \rangle (\langle c \rangle N - 1)} = \frac{4}{2(2N-1)} \approx \frac{1}{N}$となり、

$$\bar{k}(g) = \frac{2}{N}\left(4 - \frac{2}{N}\right)^{G-2} \tag{3.14}$$

と求められる。 N が大きくない場合でもp_cが二項分布${}_{2N}C_c \left(\frac{1}{N}\right)^c (1 - \frac{1}{N})^{2N-c}$で表されるとすれば $\langle c^2 \rangle = 6 - \frac{2}{N}, \langle c \rangle = 2$ なので $q = \frac{\langle c^2 \rangle - \langle c \rangle}{\langle c \rangle (\langle c \rangle N - 1)} = \frac{4}{2(2N-1)} = \frac{1}{N}$

となり同じ結果になる。指数の底を

$$\tilde{B} \equiv 4 - \frac{2}{N} \tag{3.15}$$

とすると累積祖先ループ数は

$$\bar{K}(g) = \sum_{g'=2}^{g} \bar{k}(g) \tag{3.16}$$

$$= \frac{2}{N} \frac{\tilde{B}^{g-1} - 1}{\tilde{B} - 1} \tag{3.17}$$

$$\approx \frac{2}{(\tilde{B} - 1)N} \tilde{B}^{g-1} \tag{3.18}$$

$$= \frac{2}{(3N-2)} \left(4 - \frac{2}{N}\right)^{g-1} \tag{3.19}$$

$$\approx \frac{1}{6N} 4^g \tag{3.20}$$

となる。(3.18)式の近似は G が大きい時$\tilde{B}^{G-1}$に比べて−1 を無視した。(3.19)式より N = 1, 2, 3, ･･･で$\tilde{B} = 2, 3, 10/3, \cdots$となり、N が十分大きいと 4 に漸近する。(3.20)式の近似は、分母と分子に(4-2/N)をかけ、N が大きいとすると(4-2/N)→4 とした。

3.3 Derrida モデルの実測

本節では Derrida モデルを用いて実際に累積祖先ループ数と探索世代の依存性を測定し、それと理論値を比較することで、累積祖先ループ数の式と一致するかを調べる。

3.3.1 戸籍情報

祖先ループ数を測定するために、作成された家系図はデータとして保存する必要がある。そこで家系図の親と子の関係をデータとして保存しなければならない。以下このデータを戸籍情報と呼ぶ。その内容を表 3.1 に示した。

ID	通し番号。初めの番号は 0
生年	生まれた年（Derrida モデルは世代 G）
名前	個体名
性別	雌が 0 、雄が 1
父名	父の名、存在しない場合、不明の場合 -1
母名	母の名、存在しない場合、不明の場合 -1
子数	子の数

表 3.1　戸籍情報の内容。

図 3.2 は単純 Derrida モデルによる家系図の一例であり、表 3.2 はその家系図の戸籍情報である。表 3.2 の戸籍情報から図 3.2 の家系図を作成することもできる。例えば図 3.2 の 10 と書かれた個体は名前が 10、生年 G＝1、性別は雌である。今回、Derrida モデルで作成した家系図で個体名は ID と同じ番号としている。例えば、名前が 10 の個体の ID は 10 である。次に親子関係を見ると ID 10 の父は ID5 で、母は ID0 である。ID 10 の子数は 3 であるとわかる。これらは

表 3.2 と対応している。このように家系図と戸籍情報は対応しているのでどちらか一方がわかれば他方も復元できる。以下、この戸籍情報を用いて各個体の祖先ループを調べる。

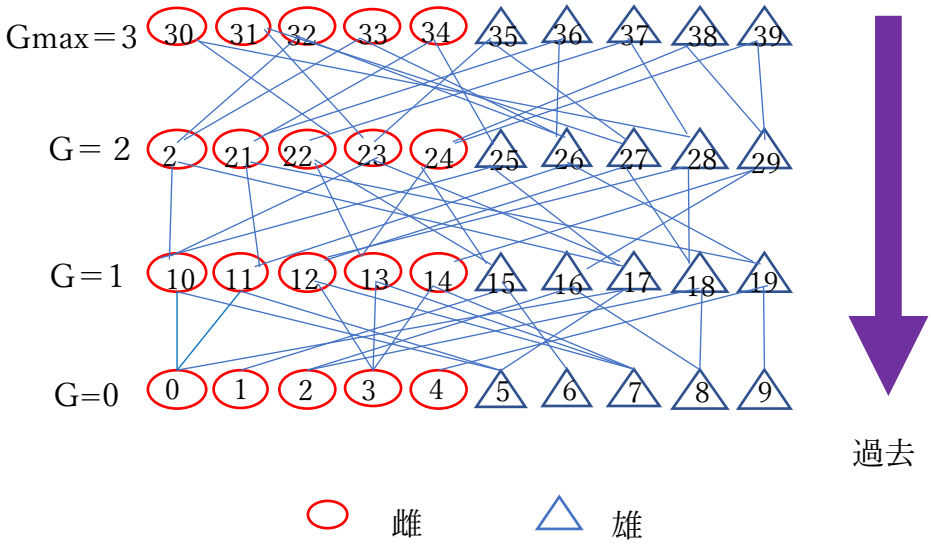

図 3.2　Derrida モデルの例。

ID	生年(G)	名前	性別	父名	母名	子数
0	0	0	0	-1	-1	3
1	0	1	0	-1	-1	2
:	:	:	:	:	:	:
10	1	10	0	5	0	3
:	:	:	:	:	:	:
23	2	23	0	17	10	1
:	:	:	:	:	:	:
38	3	38	1	29	24	0
39	3	39	1	29	24	0

表 3.2　図 3.2 の家系図の戸籍情報。

3.3.2　測定方法

累積祖先ループ数を測定するアルゴリズムは以下の通りである。戸籍情報、探索する世代数、測定個体（α個体）を必要とする。初めに戸籍情報の中から、α個体の親を探し、その親の戸籍情報から親の親を探す。この手続きを探索する世代数まで遡って繰り返すことで祖先リストを作成する(表 3.3)。祖先リストの構成は祖先番号、個体名、役割重複回数からなる。祖先番号はα個体の祖先を並べるときにその役割に応じてつけた番号であり、以下のように決める。着目する個体（α個体）の祖先番号は 1 とする。ある個体の祖先番号が j の時、その個体の母親の祖先番号を 2j、父親の祖先番号を 2j＋1 とする。するとαの母の祖先番号は 2、αの父の祖先番号は 3，αの母の母、母の父、父の母、父の父の祖先番号はそれぞれ、4、5、6、7 となる。このように祖先番号が与えられれば、α個体との続柄がわかる。役割が重複した回数を役割重複回数と呼ぶ。表 3.3 の祖先番号 16 と 30 では Z が祖先の役割を二回担っているので Z の役割重複回数は 2 回である。役割重複回数はα個体から遡る経路の数と対応する。役割重複回数が 2 回以上の個体を役割重複個体と呼ぶ。

次に祖先リストからすべての役割重複個体（t≧2）を探し、α個体とその個体までの経路のすべての組みあわせが祖先ループであるかどうかを確認する。具体的には役割重複個体までの経路の組み合わせの中で同一個体のない経路の組み合わせを数える。経路上の個体が同一個体かどうかは戸籍番号を用いて確認する。

祖先ループが複数ある例を図 3.3 に示す。着目個体αの祖先ループは黄色と水色のペア、黄色と橙色のペア、紫色と水色のペア、紫色と橙色のペアの 4 つある。このようにγ個体の子数が少ない場合でも経路の中で役割重複個体がいると、γまでの経路の数が増えるので祖先ループ数は多くなる場合がある。以上のことに注意して、探索世代ごとに祖先ループ数を測定し、探索世代と祖先ループ数の関係性を調べた。

祖先番号（役割）	名前	役割重複回数
1 （α個体）	A	1
2 （母）	B	1
3 （父）	C	1
4 （母の母）	D	1
5 （母の父）	E	1
6 （父の母）	F	1
7 （父の父）	G	1
：	：	：
16 （母の母の母の母）	Z	1
：	：	：
30 （父の父の父の母）	Z	2
：	：	：

表 3.3　祖先リストの例

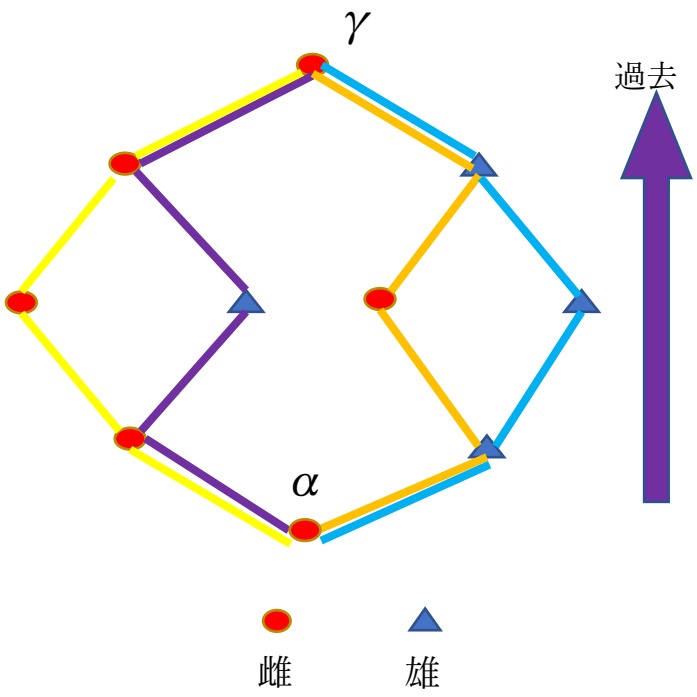

図 3.3　祖先ループが複数ある例。α が着目個体、γ が両親の共通祖先。

3.3.3　測定結果

前節の理論式によれば､指数関数の底 B は初期個体数 N に依存する｡そこで、Derrida モデルの実測でも N を変化させて、累積祖先ループ数を測定した。まず N が比較的大きい場合である。Gmax＝20 の Derrida モデルで最終世代に生まれた個体のうちランダムに一頭選んだ個体の累積祖先ループ数を測定した。ここではデータの偏りをなくすために Derrida モデルので家系図を生成するときの乱数の種を 5 通り、それぞれの家系図で着目個体を 5 通り合計 25 通りで測定した。その結果が図 3.4 である。横軸は探索世代、縦軸はその累積祖先ループ数である｡N に寄らずgが大きいところで全てのデータが4^gに漸近していることから、累積祖先ループ数は指数関数的に増大していることがわかる。また初期個体数の増加とともに累積祖先ループ数は減少している。

前節（3.20）式によれば、指数関数の底 B は N が大きいところで 4 に漸近するが、これは図 3.4 の結果と無矛盾である。(3.20)式を変形すると$6N\overline{K}(g) = 4^g$となり$6N\overline{K}(g)$は個体数 N によらないと考えられる。実際に図 3.5 を見ると、すべてのデータが4^gに重なっていることがわかる。さらに変形すると個体数 N と探索世代 g が大きなところでは比$6N\overline{K}(g)/4^g$は 1 に漸近すると考えられる。実際に図 3.6 を見ると、個体数 N と探索世代 g が大きなところで 1 に漸近していることがわかる。

次に、初期個体数 N が小さいところでの B のN依存性を確かめる。Nを 1 から 100 の間で変化させて家系図を作成し、累積祖先ループ数を測定した。その結果が図 3.7 である。横軸は探索世代、縦軸はその累積祖先ループ数である。N が小さい場合、指数の底は 4 からずれることが確認された。図 3.8 は累積祖先ループ数を指数関数と近似したときの底の値の N 依存性を表したグラフである。横軸が初期個体数、縦軸は累積祖先ループ数の指数の底である。緑線は理論値を表したグラフである。N が小さい場合は（3.20）式は使えず、(3.10）式まで戻る必要がある。(3.10) 式の子数の二乗平均$\langle c^2 \rangle$は数値的に得られたものを用いてる。結果の偏りをなくすために 10 通りの家系図を使用して計算した。黒のグラフは累積祖先ループ数の測定データの 50 通りにそれぞれフィッティングしたものの平均値である。揺らぎの中に理論値は収まっているので測定値と理論値が無矛盾であることがわかった。

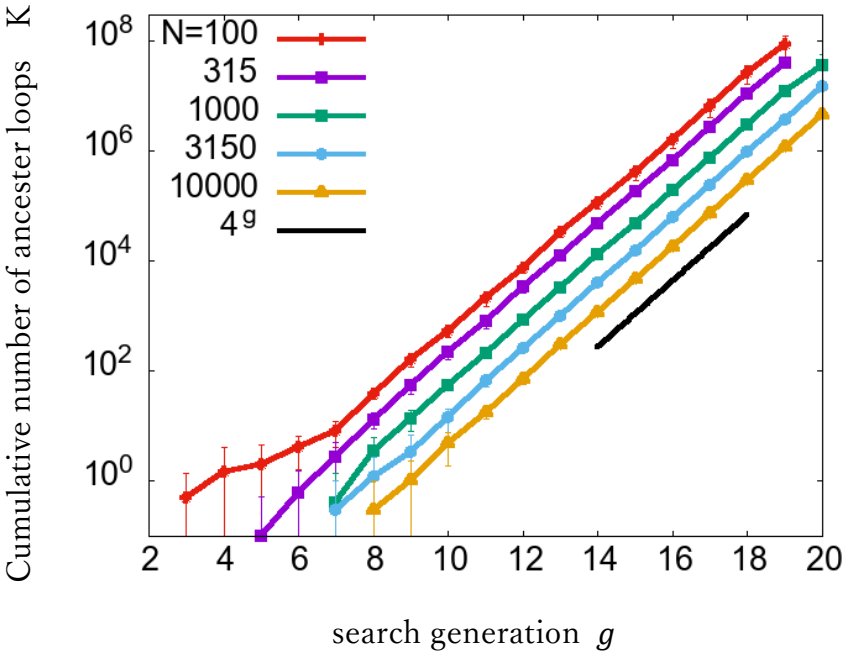

図 3.4　Derrida モデルにおける累積祖先ループ数の探索世代依存性。着目個体 5 個体の平均と標準偏差を表示。黒線は4^g。N によらず、G が大きいところで4^gに漸近している。また N の増大とともに K は減少している

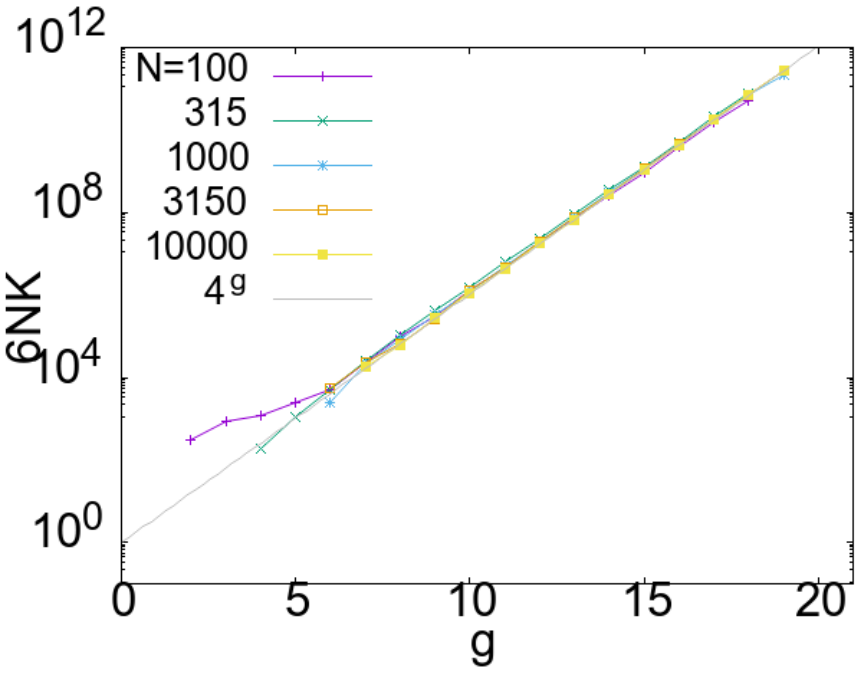

図 3.5　Derrida モデルにおける累積祖先ループ数と指数関数の関係性。着目個体 5 個体の平均を表示。6NK は N によらず 4^g に漸近している。

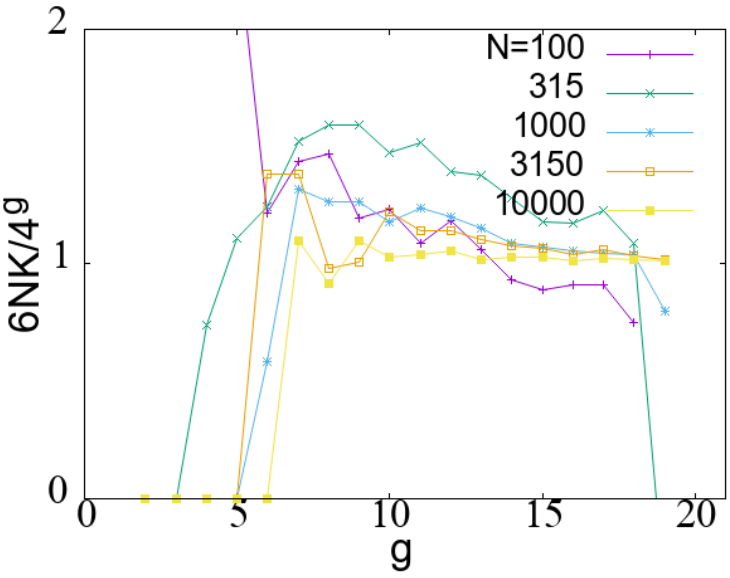

図 3.6　Derrida モデルにおける累積祖先ループ数と指数関数の関係性。着目個体 5 個体の平均を表示。$6NK/4^g$ は 1 に漸近している。

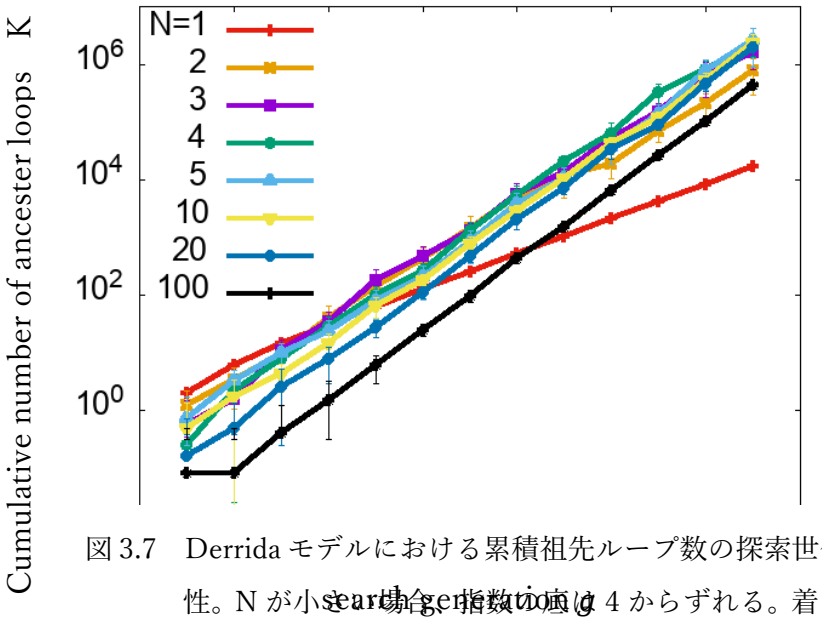

図 3.7　Derrida モデルにおける累積祖先ループ数の探索世代依存性。N が小さい場合、指数の底 g は 4 からずれる。着目個体 10 個体の平均と標準偏差を表示。

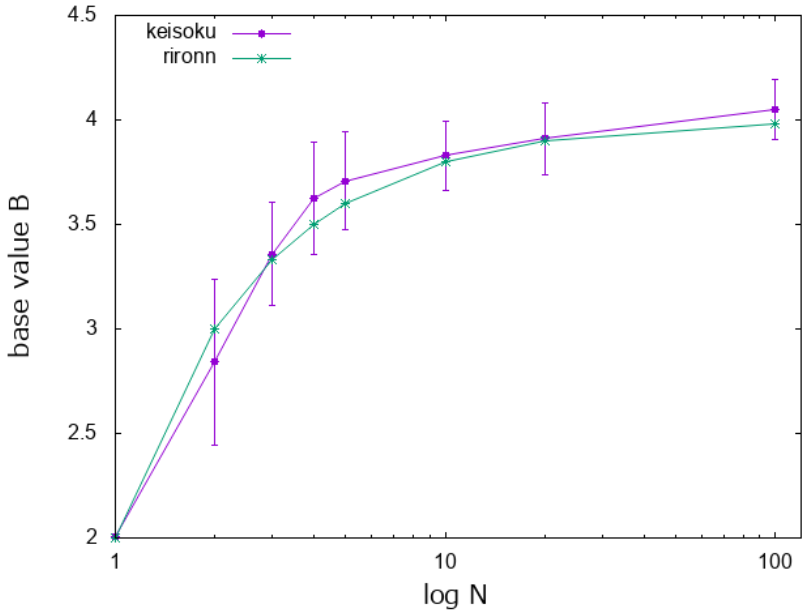

図 3.8　累積祖先ループ数を指数関数と近似したときの底の値の初期個体数依存性。測定データ 50 通りに関する平均値と標準偏差を示している。

4章　実データ

前章で用いた Derrida モデルと実際の家系図を比較するために、実際の競走馬の家系図に関する解析も行った。本章は競走馬の家系図データの説明と、解析結果を紹介し、仮想的な家系図と比較する。

4.1 競走馬の家系図

実際に家系図を長く遡ることができる生物として、競走馬に着目した。本研究で用いた競走馬のデータは 、Ikuta が競走馬の家系に関する情報を収集しているサイト pedigreequery.com で「GODOLPHIN ARABIAN」の子孫を調べたものである[4]。本研究ではその中の 1724 年から 1960 年までの家系図データを使用した。登録個体は 333461 頭である。家系図データには戸籍情報（名前、誕生年、性別、父名、母名、子数）があり、例えば HOTDISH という馬は HOT DISH、1960、M、GRATITUDE、REDHOTPOKER、6 とデータにある。

具体的に HOTDISH の家系図を見てみよう。pedigreequery.com の中で HOTDISH の家系図を 5 世代前まで検索した結果が図 4.1 である。この図から親子関係をたどることができる。HOTDISH の父の祖先と母の祖先を 4 世代遡ると次の事がわかる。HOTDISH を α とすると、BAHARAM は α の父の母の母の父であり α の母の父の父でもあることがわかる。さらに遡ると SWYNHORD、THE TETRARK、BLANFORD、FRAIAR'S DAUGHTER、BLANCHE、FRAIAR MARCUS、GARRON LASS も HOTDISH の父と母の共通祖先であるとわかる。図 4.1 をもとに図 4.2 のような祖先のネットワークを書き表すことができる。図 4.2 より HOTDISH とその父と母、父と母の共通祖先 BAHARAM を通る経路が祖先ループであることがわかる。さらに遡ると SWYNHORD、THE TETRARK も HOTDISH の祖先ループを閉じさせる祖先個体である。

実際の家系図と Derrida モデルの家系図とはいくつかの点で異なっている。詳細については 4.3 節で述べるが、ここで祖先ループに関する構造の違いについて述べておこう。HOTDISH から始まり BAHARAM で閉じる祖先ループを例に考えると、HOTDISH から BAHARAM に至る経路は、父経由だと 4 世代前だが、母経由において 3 世代前であり長さが異なる。このように、実際の家系図において祖先ループは「ちょうど G 世代で閉じる」とは限らない。しかし、G

世代までで閉じる祖先ループならば、はっきりと定義できる。我々が、祖先ループを数える場合に、累積祖先ループ数に着目したのは、実際の家系図と比較するためである。

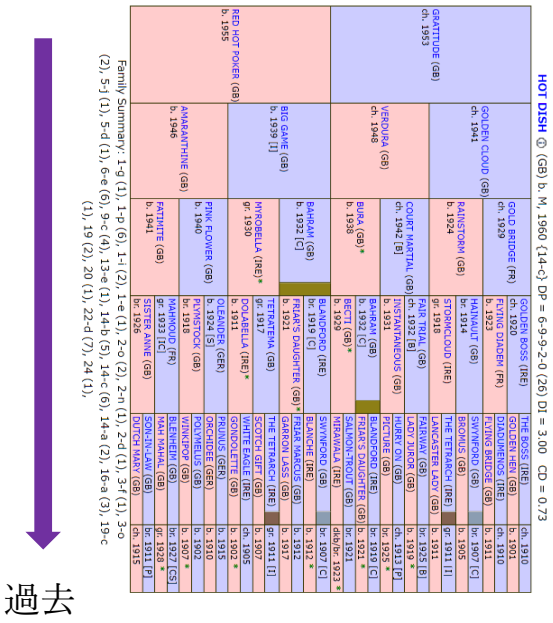

図 4.1 家系図の例。5 世代前までの全祖先を表示。水色が雄でピンクが雌を表している。pedigreequery.com より転載。

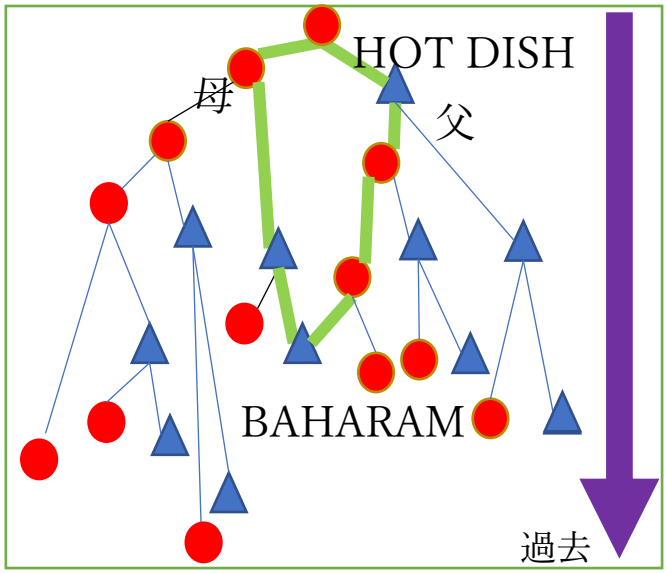

図 4.2 図 4.1 の家系図で同じ個体を同じ点で表示。4 世代前まで表示。BAHARAM は父と母の共通祖先。緑線が祖先ループ。

4.2 解析結果

Derrida モデルで作製した家系図では、累積祖先ループ数は探索世代に応じて指数関数的に増大していた。競走馬の家系図ではどのように変化するかを調べるため、α 個体の母集団として 1960 年生まれの個体 10 頭をとりそれらの累積祖先ループ数の平均を測定した。その結果が図 4.3 である。オレンジ色のグラフは図 3.4 と同様 N＝10000 の Derrida モデルの累積祖先ループ数の測定結果である。モデルにおける初期個体数は、モデルと競走馬で総個体数がおよそ等しくなることを条件として決めている。競走馬の累積祖先ループ数は Derrida モデルと同様に指数関数的に増大していることがわかる。また 2 つのグラフを指数関数$\overline{K} = mB^g$でフィッティングすると、

$$\text{モデル：B} = 4.0 \pm 0.014、m = 10^{-5.4\pm0.021} \tag{4.1}$$

$$\text{競走馬：B} = 4.1 \pm 0.063、m = 10^{-3.5\pm0.081} \tag{4.2}$$

となった。競走馬の累積祖先ループ数は Derrida モデルと同様に、およそ 4 の指数で増えていることがわかる。残念ながら競走馬の家系図に対して、理論式は適用できない。これは集団全体で世代が同期していることを仮定しているからである。また、累積祖先ループ数は同じ総個体数の Derrida モデルとくらべ多いことがわかる。この違いは競走馬の家系図と Derrida モデルと家系図の構造上の違いに起因すると考えられる。

図 4.3 で測定した競走馬の着目個体 10 頭で g＝15 までの役割重複回数と祖先ループ数の関係性を調べた。結果が図 4.4 である。横軸が α 個体と役割重複個体までの経路のすべての組みあわせの数、縦軸はその個体で閉じる祖先ループ数である。役割重複回数とともに祖先ループ数は増えていき、ある一定の割合の範囲で祖先ループが存在することがわかる。理論的な解釈は出来てきておらず、今後の課題である。

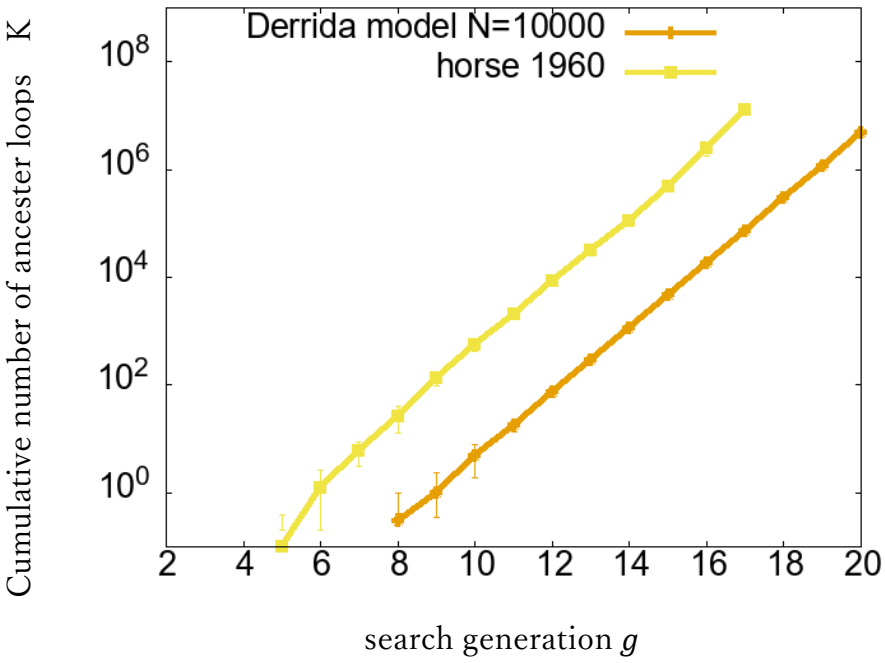

図 4.3　競走馬と Derrida モデルにおける累積祖先ループの探索世代依存性。競走馬は 1960 年生まれの 10 個体の平均と標準偏差。Derrida モデルは N=10000。

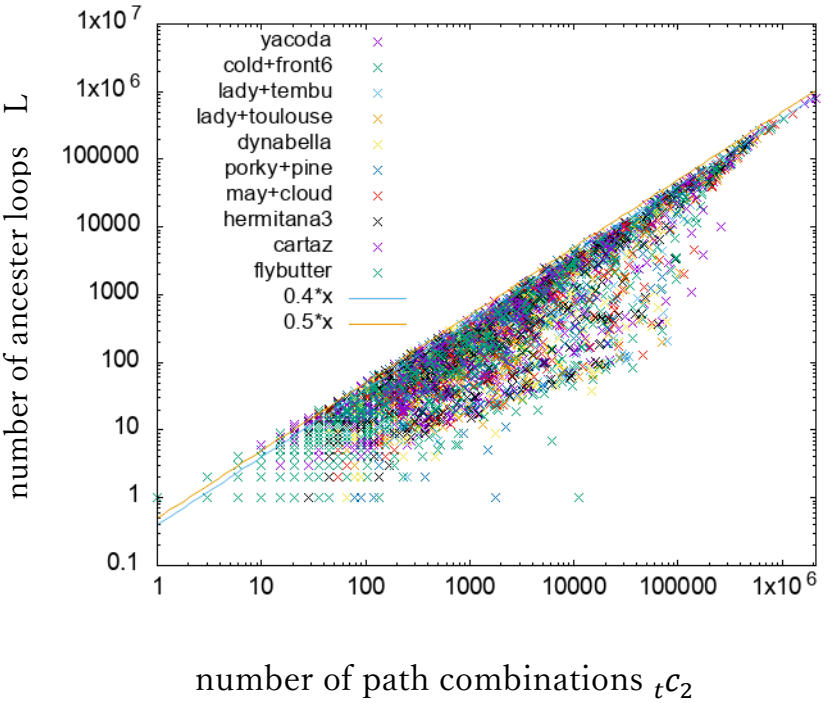

図 4.4　競走馬における祖先ループ数と役割重複回数の関係性。α 個体は図 4.3 で測定した 1960 年生まれの 10 個体。

4.3 競走馬と Derrida モデルの違い

競走馬と Derrida モデルの家系図の構造および特徴量の違いはいくつか存在する。それらについて紹介する。

4.3.1　世代の非同期性と世代間隔

Derrida モデルでは世代を G=0,1,2,…と定義できたが、競走馬の個体の生年は様々であることから、世代は同期していない。しかし親子の年齢差を見積もることで、競走馬の家系図がおよそ何世代の家系図からできているのかを見積もることができる。まず親子年齢差を調べるため、1960 年に生まれた個体を母集団として親子年齢差の確率分布を測定した。その結果が図 4.5 である。横軸が親子の年齢差で縦軸がその年齢差における確率である。親子年齢差の最頻値は 10 年で、平均値$\langle a_d \rangle = 11.2 \pm 4.3$年である。比較のため λ＝11.2 の Poisson 分布も図に示したが、やや違う形をしている。またほかの年代に生まれた個体を母集団として同様に親子年齢差を測定した。その結果が図 4.6 である。1760 年～1960 年で 20 年ごとに測定したもので、どの年代の親子年齢差の平均も 10~15 で収まっている。

最も母集団の数が多い 1960 年に生まれた個体の親子年齢差の平均値を用いて、競走馬の家系図の世代の見積りを行った。例えば 20 世代前は約 11.2×20＝224 年前である。1960 年から 20 世代前だと 1736 年になる。戸籍情報によるとこの家系図の始祖である GODLPHIARABIAN の生年は不明だが、次に古い個体の生年は 1738 年であるため約 20 世代の家系図でできていると見積もることができる。

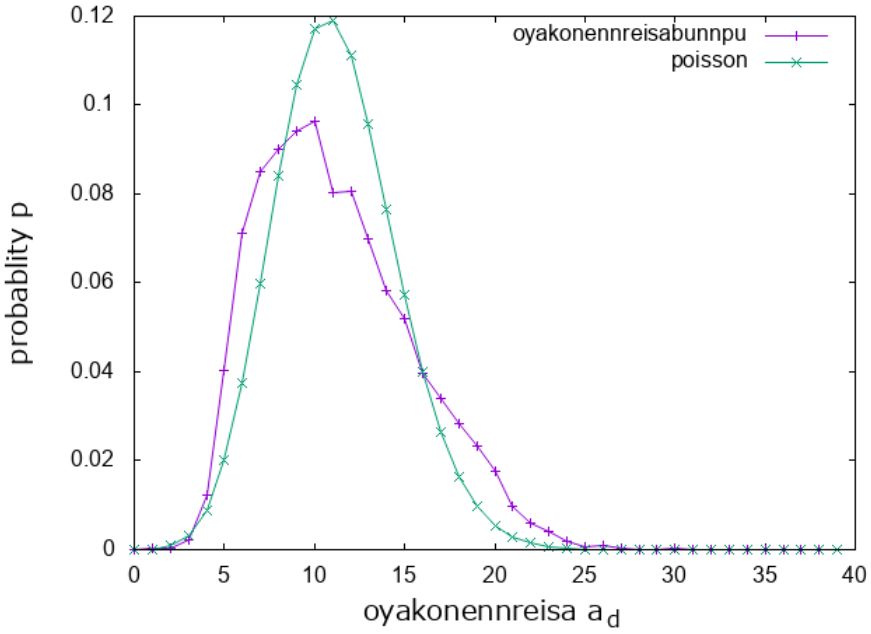

図 4.5　親子年齢差の分布。競走馬の母集団は 1960 年生まれの競走馬。 $\lambda = 11.2$ の Poisson 分布を表示。

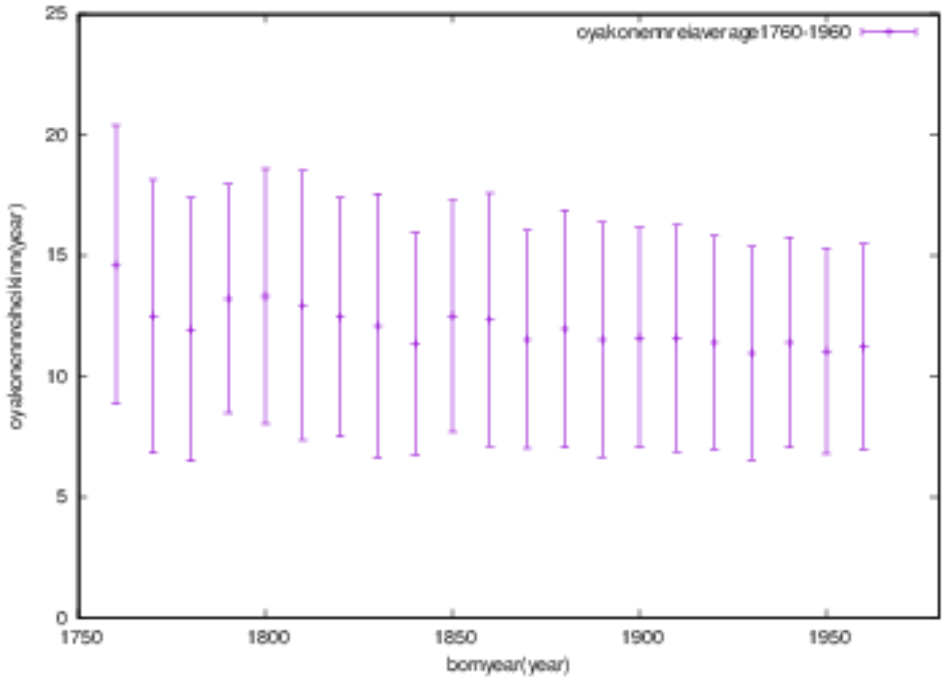

図 4.6　親子年齢差の平均値の時系列。1760 年から 1960 年の間で 20 年ごとに測定。

4.3.2 子数

競走馬と Derrida モデルの子数分布の違いを図 4.7 に示した。Derrida モデルでは、個体数が多い場合、雄も雌もほぼλ＝2 の Poisson 分布である。これに対して競走馬の子数分布は雄と雌で大きく異なる。まず雌の子数分布は Poisson 分布に近い形のグラフに見える。競走馬の雄の子数分布は雄は一頭も産まないのがおよそ半分、残り半分は子数にべき的に依存し

$$p_c^{\sigma} \approx 0.1 c^{-1.5} \tag{4.3}$$

と表すことができる。これら子数分布の違いは、生殖における雄と雌の役割の違いに起因する。雌は自分の体から子供を産むので、雄のように 500 頭産むということは考えにくい。

競走馬の生まれた年代によって子数分布が変化するのかどうかを調べた。1760 年～1960 年で 20 年ごとに測定した結果を図 4.8 に示す。どの年代でも 1～3 頭の子供を持つ馬が多いことがわかる。またこの分布データをもとに生年ごとの平均子数を計算した。その結果が図 4.9 である。年代によって平均子数の値は異なり、1850 年を除けば 2～3 付近で収まっている。また平均子数は右肩上がりになっている。原因はまだわからないが、生年が大きくなると個体数が多くなるため初期の年代と比べてより正確な値が出ているのかもしれない。1850 年のところで急激に増加していることがわかる。人為的に交配を増やすようにしたのか、それとも偶然大きな値を示すだけなのか原因はわからないが、他の生物の家系図でも同様の事があるのか調べてみる必要がある。

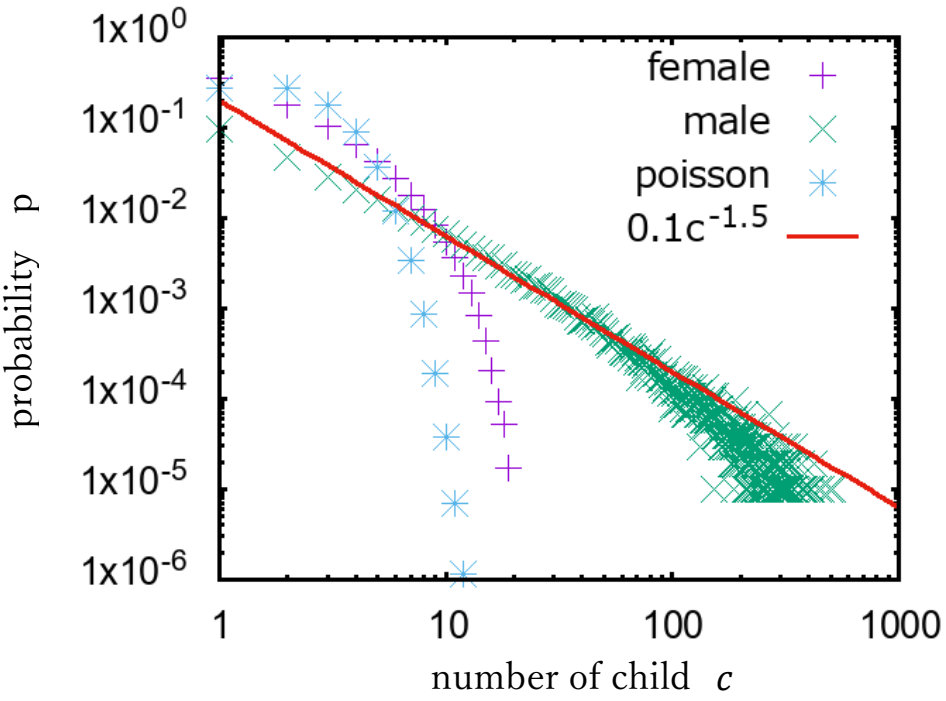

図 4.7 競走馬の家系図と Derrida モデルの子数分布。
Derrida モデルは λ =2 の Poisson 分布を表示。

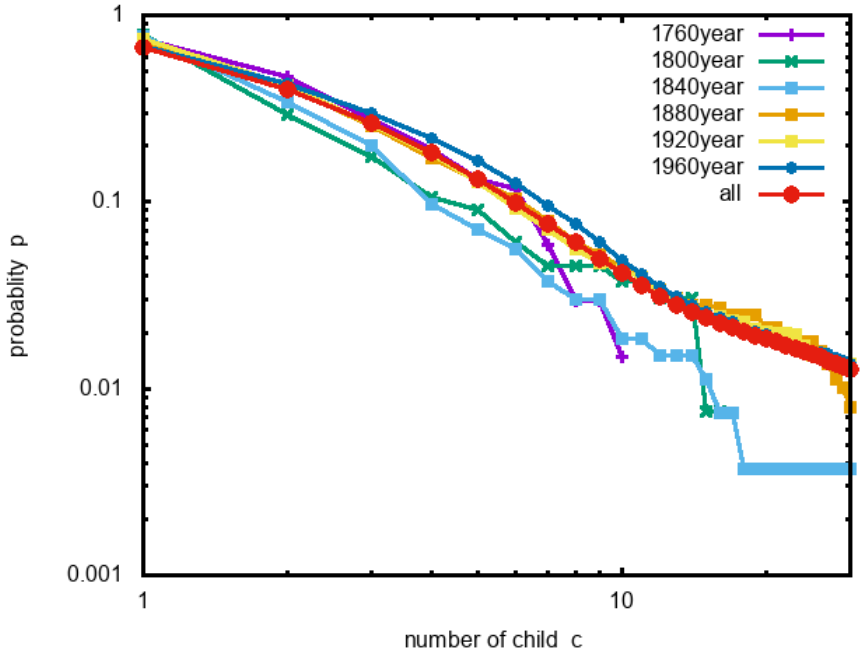

図 4.8　競走馬の子数分布の年代依存性

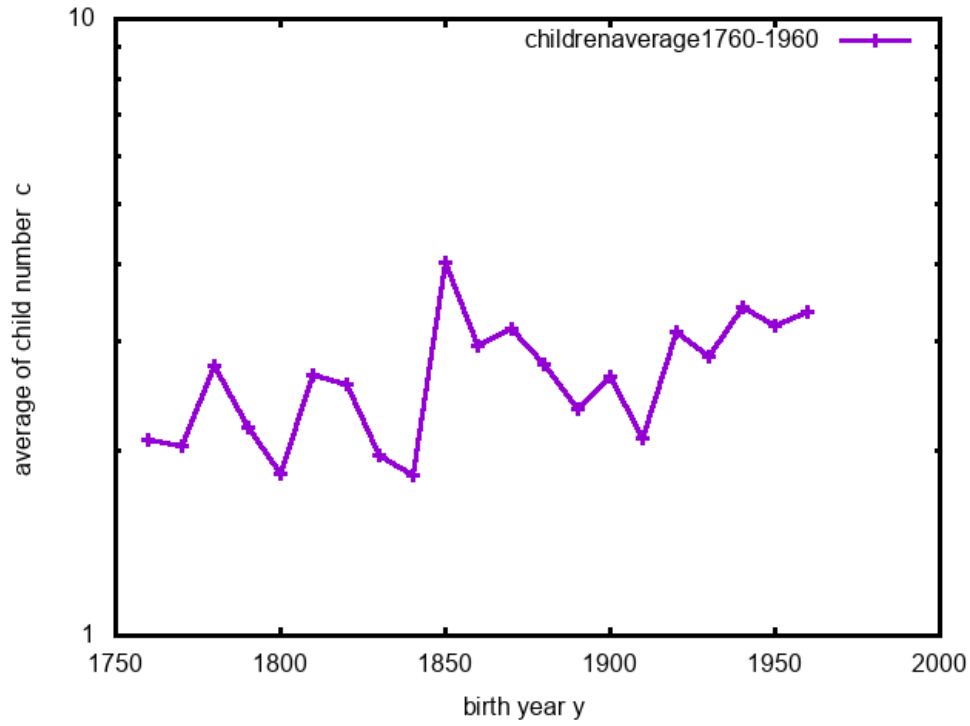

図 4.9　競走馬の子数平均の時系列。1760 年から 1960 年の間で 10 年ごとに測定。

4.3.3 個体数

競走馬の家系図と Derrida モデルの家系図では個体数の比と増加率に違いがある。まず雄と雌の個体数の比の違いであるが、これは図 4.10 と図 4.11 で示した。図 4.10 で雄と雌の個体数のグラフが重なっていることからわかるように Derrida モデルでは個体数の比はおよそ 1：1 である。また個体数はやや減少しているが、これは Derrida モデルの実装に際して用いた条件「ペアを組む数は個体数が少ない方の性別にの数に合わせ条件（A.5.1)」と「ペアを組めなかった個体は子供を産まないものとする（A5.2)」に起因していると考えられる。一方、競走馬の場合、図 4.11 で雄と雌の個体数のグラフがおよそ平行であることから、雌と雄の個体数の比はおよそ 2.5:1 という一定の割合を保っていることがわかる。なお、雄と雌の総個体数は指数関数的に増加しており、西暦を y、y 年生まれの個体数を$N_{all}(y)$とすると、

$$N_{all}(y) \fallingdotseq 10^{(0.0167y-28.5)} \tag{4.4}$$

と近似することができる。

次に個体数の世代ごとの増加率を二種類の方法で見積もり比較した。第一の方法は、ある世代の総個体数Nとその親世代の総個体数N'の比N/N'から見積もる方法であり、これを$\rho = N/N'$とする。これに対して、第二の方法は、親の世代が生んだ子の総数B'と親の世代の総個体数N'の比B'/N'から見積もる方法であり、これを$C = B/2N'$とする。右辺の分母の 2 は親が二頭必要なことを意味している。Derrida モデルであれば、$\rho = C$となるが、競走馬の家系図では。親子年齢差も分布しており、親が不明な個体も存在するため、一致するとは限らない。ここでは 1960 年に生まれた個体の親に着目し実測した結果を報告する。以下、親子年齢差をδとし、その分布は実測された分布$f(\delta)$（図 4.5）を用いる。 y 年生まれの個体の総数を$N_{all}(y)$を用いると、第一の方法より、

$$N_{all}(1960) = \sum_{\delta} f(\delta) \times \rho(1960 - \delta) \times N_{all}(1960 - \delta) \tag{4.5}$$

が成り立つ。ここで$\rho(1960 - \delta)$は 1960 年よりδ年前の個体数からの増加率である$\rho(1960 - \delta) \approx \bar{\rho}(1960)$と近似して$\bar{\rho}(1960)$について解くと、

$$\bar{\rho}(1960) = \frac{N_{all}(1960)}{\sum_{\delta} f(\delta) \times N_{all}(1960 - \delta)} \tag{4.6}$$

となる。

第二の方法では予測値$\bar{C}(1960)$は 1960 年の個体の親が一頭当たり産む子供の数で、1960 年の 1 世代前に生まれた個体が産んだ子の数$C(1960 - \delta)$と f(δ)をもとに求めた値であり

$$\bar{C}(1960) = \sum_{\delta} f(\delta) \times C(1960 - \delta) \tag{4.7}$$

$$= \frac{1}{2} \times \sum_{\delta} f(\delta) \times \frac{B_{all}(1960 - \delta)}{N_{all}(1960 - \delta)} \tag{4.8}$$

と表す。$C(1960 - \delta)$は$1960 - \delta$年に生まれた個体の総数$N_{all}(1960 - \delta)$と$1960 - \delta$年に生まれた個体が産んだ子の総数$B_{all}(1960 - \delta)$を用いて計算した値である。計算結果は$\bar{\rho}(1960) = 1.74, \bar{C}(1960)$=1.68 となった。両者の値は少しずれはあるが、1 世代でおよそ 1.7 倍増加すると見積もることができる。

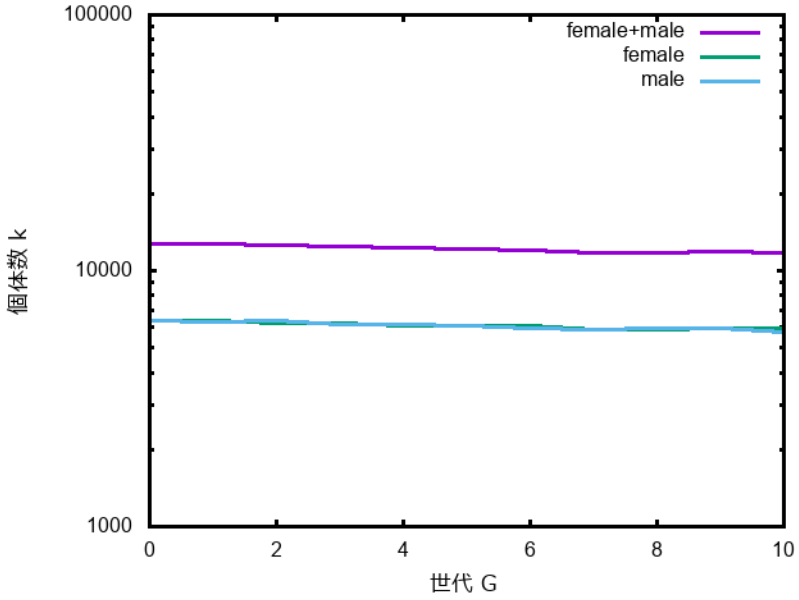

図 4.10　Derrida モデル（n=6400）の個体数の時系列

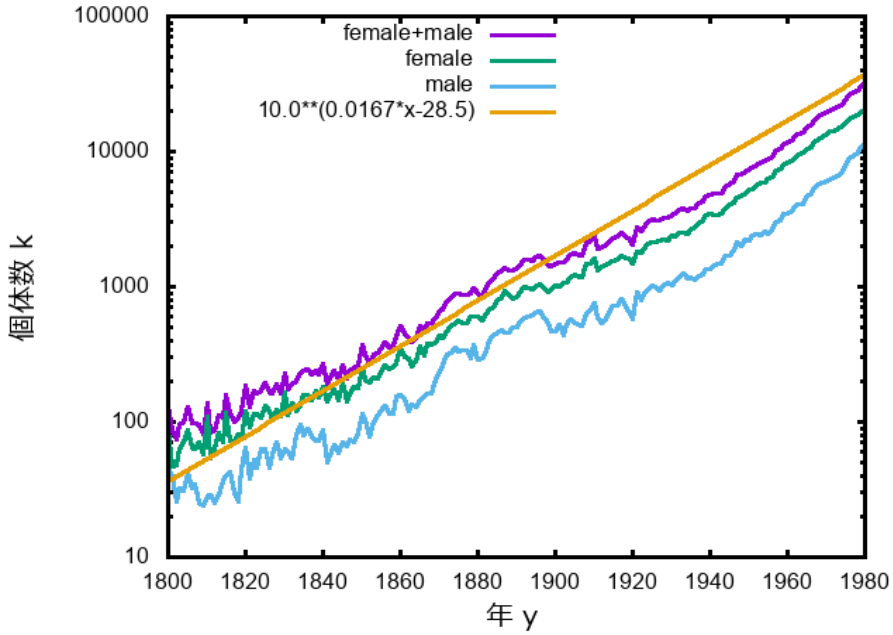

図 4.11 競走馬の個体数の時系列

4.3.4　配偶制度

Derrida モデルは一夫一妻制である。一方、競走馬は少なくとも一夫一妻制ではない。その例を二、三紹介する。表 4.1 は SISTERTOJUNO の子の戸籍情報である。子供を産む際のペアの名前が少なくと 3 種類以上あることがわかる。さらに表にある FLORIZEL が親である馬の戸籍情報も調べた。そのリストが表 4.2 である。表は一部であるが少なくともペアが 4 種類以上あることがわかる。乱婚制に近いような配偶制度だと考えられる。

個体名	生年	性別	父名	母名	子数
DARLING40	1775	0	ANTINOUS2	SISTERTOJUNO	0
ADMIRAL17	1779	1	FLORIZEl	SISTERTOJUNO	0
FAME2	1776	0	PANTALOON2	SISTERTOJUNO	2
DIOMED	1777	1	FLORIZEL	SISTERTOJUNO	103

表 4.1　SISTERTOJUNO を親に持つ馬の戸籍情報。

NIMBLE5	1784	0	FLORIZEL	RANTIPOLE	5
ULYSEES8	1777	1	FLORIZEL	SPRITE	2
PINDAR	1781	1	FLORIZEL	THALIA	0
SIRBUNBURYSFILLY	1778	0	FLORIZEL	SULATANA	0

表 4.2　FLORIZEL を親に持つ馬の戸籍情報。

4.3.5 不明個体の存在

Derrida モデルでは初期個体を除いて、親子関係は全て与えられる。これに対して、競走馬の場合、個体がデータベースに載っていない場合がある。またある個体の子がすべて載っているとも限らない。さらにまたデータベースに間

違いがあることも考えられる。これらは、実データを扱う上で避けては通れないものである。

個体から祖先方向に辿る際に、不明な個体があればそこから先の構造を正しく解析することができない。例えば、着目個体の父親が不明な場合、その個体の祖先ループは測定できない。このようなケースを避けるため、家系図をある程度まで遡ることができるような個体のみに着目している。

具体的には着目個体に対して祖先不明率を定義し、測定することで着目個体の家系図の確からしさを定量化した。祖先不明率r_uは確認されている先祖個体Aとその個体までの探索世代数g_Aを用いて

$$r_u \equiv 1 - \sum_A 2^{-2g_A} \tag{4.9}$$

で定義される。すべての先祖個体が確認される場合$r_u = 0$、逆にすべての先祖が不明な場合$r_u = 1$となる。例えば着目の個体の父側の先祖個体がすべて確認されているが、母が不明の場合は祖先不明率は 0.5 になる。本研究で測定した競走馬 10 頭の家系図の祖先不明率を表したのが表 4.3 である。祖先不明率の最大値は CARTRAZO という馬の家系図で$r_u = 0.0143$である。

個体名	祖先不明率
YACODA	0.000016
COLDFRONT6	0.000016
LADYTEMBU	0.000015
LADYTOULOUSE	0.000016
DYNABELLA	0.000016
PORKYPINE	0.000016
MAYCLOUD	0.000017
HERMITANA	0.000016
CARTAZO	0.000143
FLYBUTTER	0.000019

表 4.3　測定した競走馬の家系図の祖先不明率。

5 章　拡張モデル

累積祖先ループ数はどの家系図においても指数関数的に増大するのかどうかという疑問が生じた。この章では新たに作成した家系図モデルとその累積祖先ループ数に関する結果を紹介する。

5.1 一夫二妻制

一夫二妻制の家系図モデルを作製した。この家系図モデルは以下の条件を満たす。

A'.1 初期個体数は 3 N 体（雄 N 体、雌 2 N 体）とする

A'.2 各世代の総個体数は必ず 3 N になる

A'.3 各個体は 1 世代しか生きられない

A'.4 子は必ず親の次の世代で生まれる

A'.5 配偶制度は一夫二妻制とする

A'.6 配偶する相手は同じ世代の個体からランダムに 2 体選ぶ

A'.6.1 配偶の組み合わせの数は雄の数である N である。

A'.7 子の性別は雌 50%、雄 50%の確率で振り分ける

A'.8 雄と雌 2 頭の組み合わせの 1 グループで子供を産むものとする。

A'.9 雄の子数 c は λ=3 の Poisson 分布に従う。雌は以下の操作で子数が決まる。

A'.9.1 生まれた子供の母は一頭ずつ交互に振り分けて決める。

A'.9.2 子供の世代の個体数の総数がちょうど 3 N になるまで乱数を振り直す。

これらの条件をもとに以下のプロセスで仮想的な家系図を作製した。

B'.1 G＝0 世代の個体を雌 N 体、雄 2N 体用意する

B'.2 上記の配偶制度の条件 A'.6、A'.6.1 をもとにペアを組む

B'.3 ペアの数分 Poisson 分布に従った乱数を振り、子数を決定し
性別を振り分ける

B'.4 Gmax 世代まで同様に繰り返す

上記の条件とプロセスをもとに作成した家系図の例が図 5.1 である。一夫二妻制の家系図においては雄と雌で子数分布に違いがある。雄の子数分布は個体数が大きい時に Poisson 分布に漸近するが、雌の子数は実装する際の条件 A'.7.1 が加わるので Poisson 分布ではないことがわかる。雄と雌の子数分布を測定した結果を図 5.2 と図 5.3 に示す。図 5.2 は N＝100 の家系図の子数確率密度関数を示している。雄はλ＝3 の Poisson 分布に近いが、雌は半数以上が 1 か 2 の子数をもつような分布になっていることがわかる。解析対象となる家系図すべての子数分布が図 5.3 である。N＝ 1 の家系図の雄の子数の確率密度分布は N＞1 の場合の雄の子数分布と大きく異なるように見えるが子数平均は変わらない。

3.2 節で導いた累積祖先ループ数の理論的な見積もりは、単純 Derrida モデルを仮定していたが、個体数および子数の確率密度関数に対する雄雌の対称性が破れているような場合にも拡張できる。計算の詳細は付録にあるが、それによると

$$\overline{K}(g) = \widetilde{m}\tilde{B}^{g} \tag{5.1}$$

$$\tilde{B} \equiv 4 - \sum_{\sigma} q^{\sigma} \tag{5.2}$$

$$\widetilde{m} \equiv \frac{\sum_{\sigma} q^{\sigma}}{(3 - \sum_{\sigma} q^{\sigma})\tilde{B}} \tag{5.3}$$

が成り立つ。ここでq^{σ}は、性別σに対する停止確率であり、

$$q^{\sigma} = \frac{\langle c^2 \rangle^{\sigma} - \langle c \rangle^{\sigma}}{\langle c \rangle^{\sigma}(\langle c \rangle^{\sigma} N^{\sigma} - 1)} \tag{5.4}$$

で与えられる。一夫二妻制の家系図では、N^{σ}はパラメータであり、$\langle c \rangle^{♂} = 3$、$\langle c \rangle^{♀} = 1.5$と決まっているが、$\langle c^2 \rangle^{♂}$、$\langle c^2 \rangle^{♀}$は作成された家系図から実測したものを用いる。

図 5.4 と図 5.5 は$\langle c^2 \rangle^{♂}$、$\langle c^2 \rangle^{♀}$および$q^{♂}$、$q^{♀}$の N 依存性を表している。図 5.4 を見ると$\langle c^2 \rangle^{♂}$、$\langle c^2 \rangle^{♀}$は N の増加関数であることがわかる。平均子数は変わらないが、子数の二乗平均は N によって変化することがわかる。また Derrida モデルと異なり q の値も雄と雌で異なる値になる。Derrida モデルと同様の方法で累積祖先ループ数を実測し、理論式（5.1)、(5.2)、(5.3）と比較した。

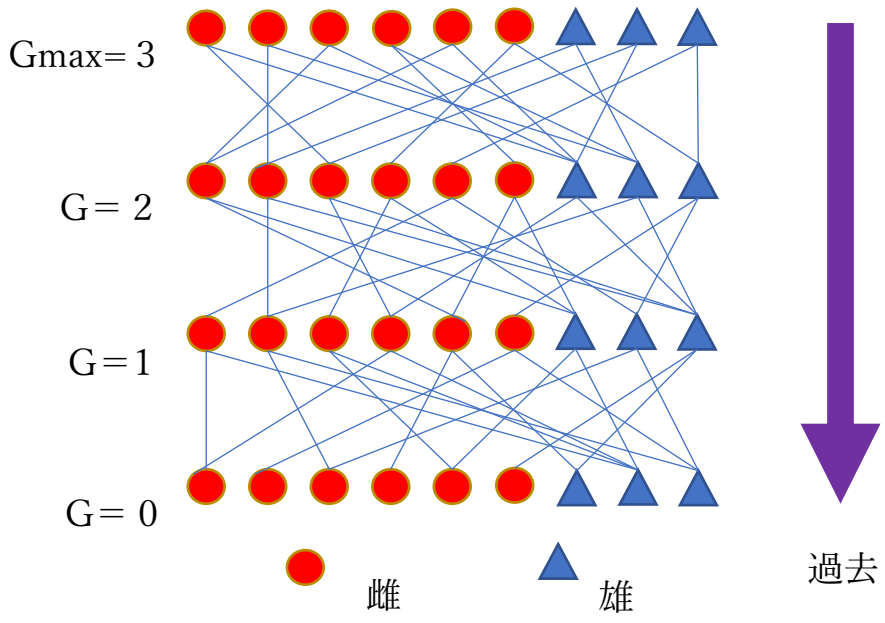

図 5.1　一夫二妻制の家系図モデルの例。

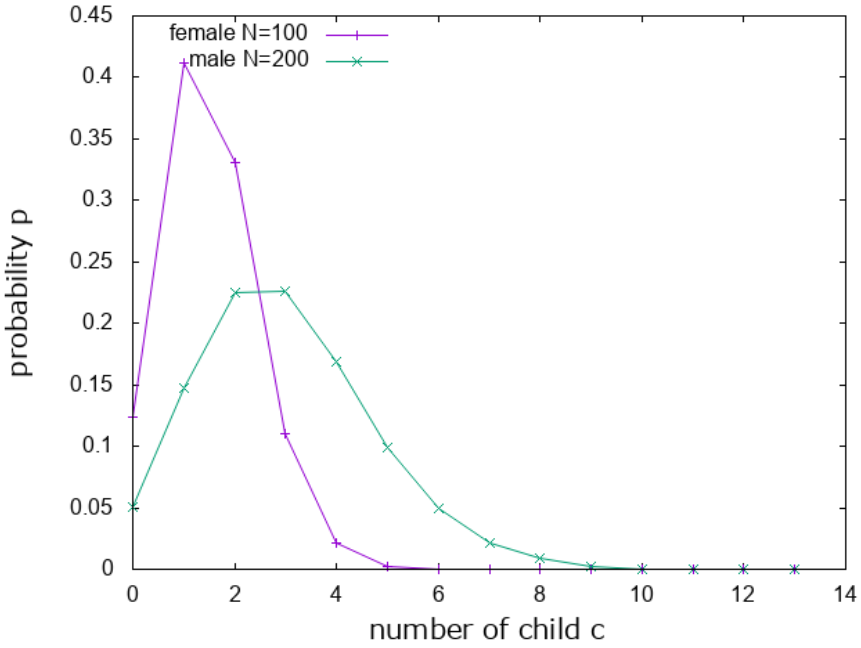

図 5.2　$N^{♂}$ =100の場合の雄と雌の子数分布。

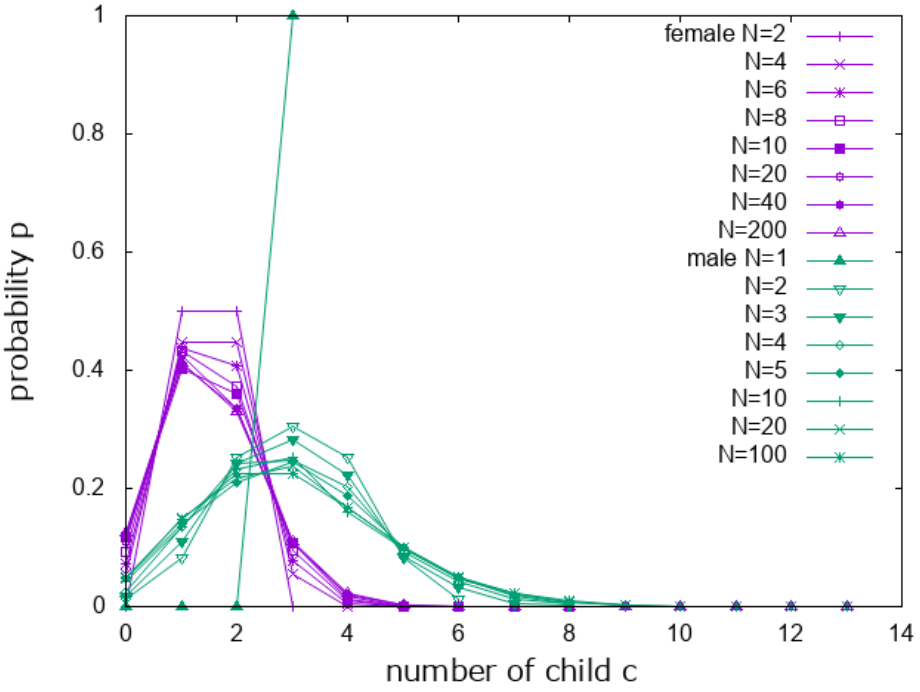

図 5.3　一夫二妻制の家系図の子数分布。

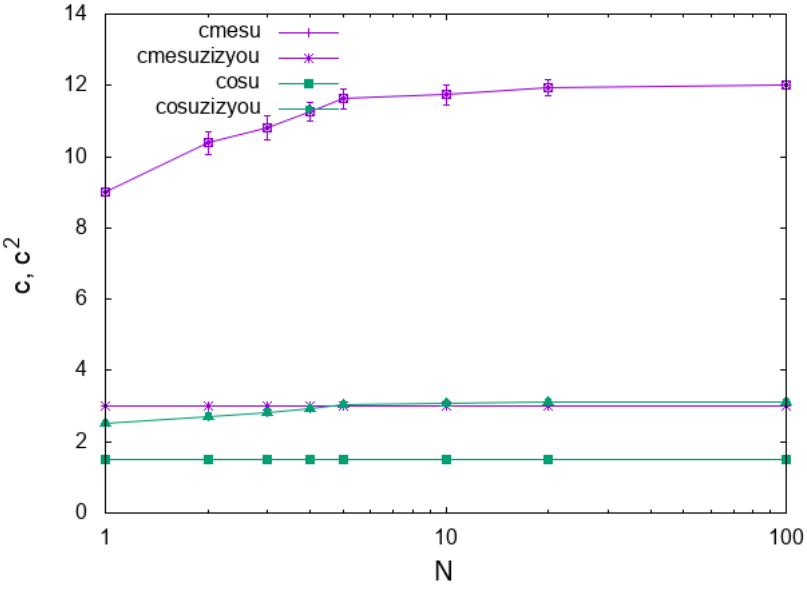

図 5.4　一夫二妻制の家系図の$\langle c \rangle, \langle c^2 \rangle$の値。

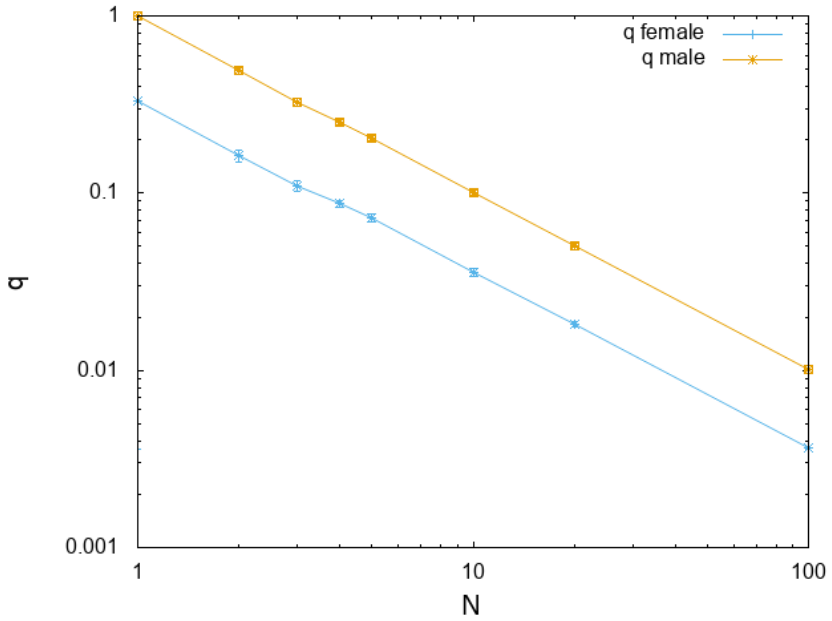

図 5.5　一夫二妻制の家系図の$q^{♀}, q^{♂}$の値。

5.2 解析結果

Gmax＝20 の Derrida モデルで最終世代に生まれた個体のうちランダムに一頭選んだ個体の累積祖先ループ数を測定した。ここではデータの偏りをなくすために家系図を生成するとき乱数の種を 10 通り、それぞれの家系図で着目個体を複数選び、測定したデータの平均と標準偏差を示している。その結果が図 5.6 である。横軸は探索世代、縦軸はその累積祖先ループ数である。図より探索世代がある程度大きくなると、累積祖先ループ数は探索世代に対して、指数関数的に増大していることがわかる。

指数の底の理論値と測定値を比較したグラフが図 5.7 である。図 5.7 の緑の実線は、異なる 10 通りの家系図と複数の着目個体それぞれに対して測定した累積祖先ループ数 K(g)をフィッティングすることで求めた底の平均と標準偏差を示している。また紫の実線は、$\langle c^2 \rangle^{\male}$、$\langle c^2 \rangle^{\female}$を測定し、それから（5.2）式で見積もられた理論値とその標準偏差を示す。Derrida モデルの家系図同様に、初期個体数が十分大きくなると 4 に収束していくことが明らかになった。測定値のゆらぎの中にと理論値が入るため、理論値と測定値は無矛盾である。

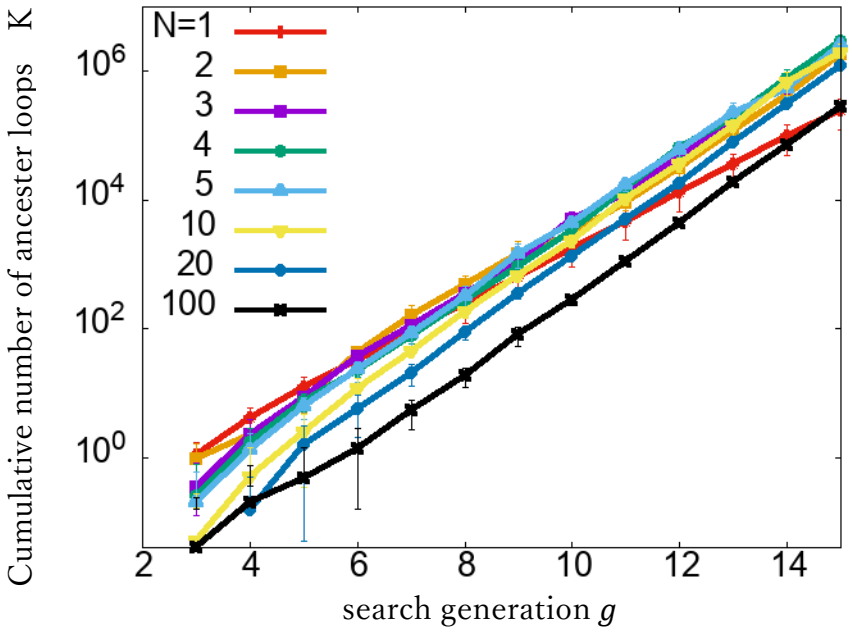

図 5.6　一夫二妻制の家系図モデルにおける累積祖先ループ数の探索世代依存性。異なる 10 通りの家系図と複数の着目個体の平均と標準偏差を表示。

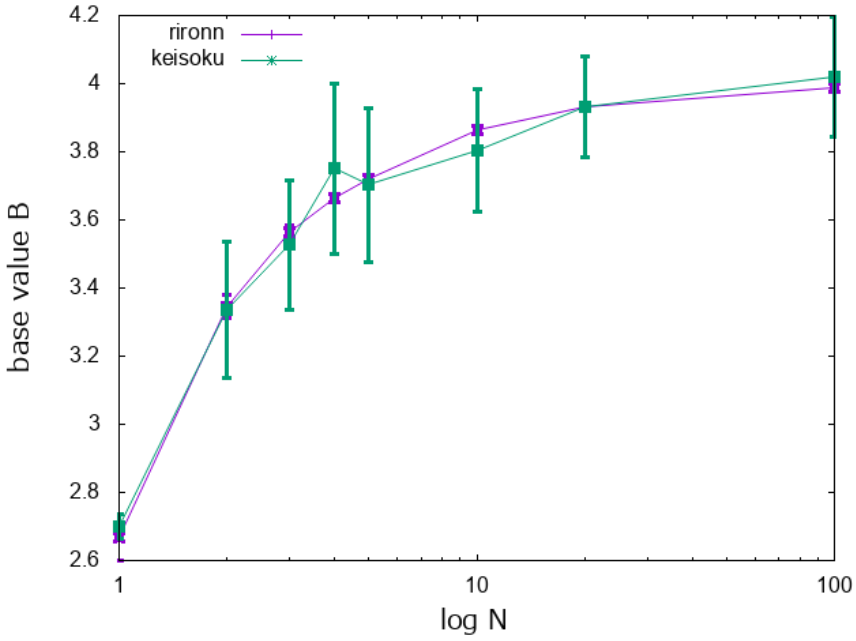

図 5.7　累積祖先ループ数を指数関数と近似したときの底の値の初期個体依存性。異なる 10 通りの家系図と複数の着目個体に関する平均値と標準偏差を示している。紫は理論値。

6 章　結論

6.1 まとめ

本研究では家系図ネットワークの複雑な構造を定量化するために、祖先ループを定義し、その数に着目した。まず、個体数が世代によらない単純 Derrida モデルによって作製された仮想的な家系図を対象とした。単純 Derrida モデルで作製される家系図の集団に対して累積祖先ループ数の探索世代依存性を理論的に見積もることに成功した。それによれば累積祖先ループ数は探索世代が大きくなると指数関数的に増大していくことがわかった。また累積祖先ループ数の指数の底は家系図の初期個体数と関係し、初期個体数が十分大きくなると 4 に収束していくことがわかった。理論の妥当性を確認するため、実際に家系図の集団を作製し累積祖先ループ数を実測したところ、理論と無矛盾な結果を得た。

次に実際の家系図として競走馬の家系図に着目し、同じ年に生まれた個体の累積祖先ループ数を測定した。その結果、競走馬も Derrida モデルと同様に探策世代が大きくなるにつれて指数関数的に増大することがわかった。また競走馬の累積祖先ループ数は総個体数が同じ程度の Derrida モデルと比べると指数の底 B は同じような値を示したが、係数mは大きな値を示した。累積祖先ループ数は家系図によらずサイズが十分に大きいと指数の底は 4 に収束するのではないかと考えられる。一方、係数 m の違いは構造の違いに起因すると考え、Derrida モデルと競走馬の家系図の構造を比較した。世代の重複の有無、雄雌の対称性の有無、子数分布、配偶制度など様々な違いがあるが、どれが係数mすなわち累積祖先ループ数に影響を与えるのかはよくわかっていない。しかし、子数分布の形（子を産まない個体の割合や、子を大量に産む個体の有無）が重要なのではないかと思われる。

さらに他の構造の家系図でも、累積祖先ループ数は指数関数的に増大するのか、また初期個体数が充分大きいと指数の底が 4 に収束するのかを調べるため、一夫二妻制に着目した。まず雄雌対称性を破るように Derrida モデルを拡張し、理論も拡張した。拡張した理論を適用した結果は、雄雌対称性が保たれている場合と同様に累積祖先ループ数は指数関数的に増大し、初期個体数が充分大きいと指数の底が 4 に収束することがわかった。実際に、一夫二妻制の家系図集団を作製し、累積祖先ループ数を実測することで、理論の妥当性を検証することがで

きた。

6.2 今後の課題

第一に、競走馬の家系図の子数分布、世代、配偶制度を定量化し、より実際の家系図に近い家系図モデルを提案し、累積祖先ループを測定することである。競走馬の家系図は世代を定義できず世代 G ではじめて閉じる条件をもとに祖先ループ数を計算するのが難しい。世代を親子年齢差の測定結果を用いて時間的に区切れば計算可能だと考えられる。第二に子数分布、世代、配偶制度などのパラメータを変更して累積祖先ループ数を測定することで、累積祖先ループに影響を及ぼす特徴量を特定することも課題である。

最後に遺伝学における近交係数との関係について触れておこう。ある個体の近交係数 F はその個体の両親の全ての共通祖先 A に対して$F = \sum_A \frac{1+F_A}{2^{n+1}}$で与えられる。この A に関する和が、本研究で求められたと祖先ループ数の数だけある。また，n は両親の親等であるが，これは祖先ループの長さに対応する。今回の計算で N が大きいとき祖先ループ数はおよそ4^gで増加するので，$\sum_A$ も4^gで増加する。世代 g で閉じる祖先ループの場合 n＝2（g—1）になる。すなわち，g が増えると，

$$F = \sum_A \frac{1+F_A}{2^{2(g-1)+1}} \approx \sum_{g=2}^{\infty} \bar{k}(g) \frac{2(1+F_A)}{4^g} \tag{6.1}$$

となる。これは、どれだけ過去へ遡っても、近交係数への寄与が有限に残ることを示唆している。これが成り立つかどうか見ることも課題である。

謝辞

本研究に取り組むにあたり、計算法のアドバイスや細部に渡りご指導いただきました水口毅准教授には、深く感謝致します。そして両親には大学院進学を支援していただきとても感謝しています。この感謝の念を忘れずにこれからも日々精進していきたいと思います。

参考文献

[1]D. J. Watts and S. H. Strogatz, Nature, **393** (1998) pp.440-442

[2] S. Gualdi, M. Medo and Y.-C. Zhang, EPL, **96** (2011) 18004 pp.1-6

[3] B. Derrida, S. C. Manrubia and D. H. Zanette, J. theor. Biol., **203** (2000) pp.303-315

[4]生田成望，大阪府立大学大学院工学研究科修士論文（2014）

[5] Pedigree Online Thoroughbred Database http://www.pedigreequery.com/

付録 A　拡張モデルにおける累積祖先ループ数に関する理論式

雄雌対称性はないが、個体数と子数分布が G によらないような場合を考える。世代 G の総個体数は

$$N^{♀} + N^{♂} = \sum_{\sigma} N^{\sigma} \tag{A.1}$$

であり、以下の式が成り立つ。

$$\sum_{\sigma} N^{\sigma} = N^{♀}\langle c\rangle^{♀} + N^{♂}\langle c\rangle^{♂} \tag{A.2}$$

したがって、性σの平均子数は

$$\langle c\rangle^{\sigma} \equiv \frac{\sum_{\sigma} N^{\sigma}}{N^{\sigma}} \tag{A.3}$$

となる。例えば $N^{♀}$:$N^{♂}$=2:1 ならば、$\langle c\rangle^{♀} = \frac{3}{2}, \langle c\rangle^{♂} = 3$である。単純 Derrida モデルと異なり、雄雌対称ではないので、世代 g で閉じる確率は式 (3.9)で g 依存性を無くした式

$$q_{\xi\eta} = \begin{cases} \dfrac{\langle c^2\rangle^{♀} - \langle c\rangle^{♀}}{\langle c\rangle^{♀}(N^{♀}(g)\langle c\rangle^{♀} - 1)} & if\ \sigma(\xi(g)) = \sigma(\eta(g)) = ♀ \\ \dfrac{\langle c^2\rangle^{♂} - \langle c\rangle^{♂}}{\langle c\rangle^{♂}(N^{♂}(g)\langle c\rangle^{♂} - 1)} & if\ \sigma(\xi(g)) = \sigma(\eta(g)) = ♂ \\ 0 & otherwise \end{cases} \tag{A.4}$$

まで戻る必要がある。1 世代で二本の経路が閉じる確率は

$$q^{\sigma} \equiv \frac{\langle c^2\rangle^{\sigma} - \langle c\rangle^{\sigma}}{\langle c\rangle^{\sigma}(N^{\sigma}(g)\langle c\rangle^{\sigma} - 1)} \tag{A.5}$$

であり、$\langle c\rangle^{\sigma}$と$\langle c^2\rangle^{\sigma}$を計算する必要がある。特に、二項分布あるいは Poisson 分布の場合、式（A.3)と$\langle c^2\rangle^{\sigma} = \langle c\rangle^{\sigma}\left(1-\frac{1}{N^{\sigma}}\right) + (\langle c\rangle^{\sigma})^2$より

$$q^{\sigma} = \frac{1}{N^{\sigma}} \qquad (A.6)$$

となる。有限サイズの数値計算では式（A.5）を実際に測定するべきである。
式（3.13)も σ によって場合分けしなければならない。これを

$$\bar{k}(g) = \sum_{\xi,\eta}\prod_{g'=2}^{g-1}\sum_{\sigma}\left(1-q^{\sigma}\delta_{\sigma\sigma\left(\xi(g')\right)}\delta_{\sigma\sigma\left(\eta(g')\right)}\right)\times q^{\sigma}\delta_{\sigma\sigma(\xi(g))}\delta_{\sigma\sigma(\eta(g))} \qquad (A.7)$$

と書こう。式（3.3)(3.5）に比べて$\sum_{\sigma}$と Kronecker のデルタが増えているが、これは σ＝♀,♂を別々に扱うことを意味している。経路ξとηの世代 g での個体の性がどちらも雄になる確率 1/4、どちらも雌になる確率 1/4 である。したがって、世代 g でξとηの祖先が同じ個体になる確率は$q^{♀}/4 + q^{♂}/4$となる。これを停止確率とする幾何分布を考えると、ちょうど世代 g で閉じる祖先ループ数の期待値は

$$\bar{k}(g) = \left(4-\sum_{\sigma}q^{\sigma}\right)^{g-2}\times\sum_{\sigma}q^{\sigma} \qquad (A.8)$$

となる。累積祖先ループ数$\bar{K}$は

$$\bar{K}(g) = \frac{\sum_{\sigma}q^{\sigma}}{3-\sum_{\sigma}q^{\sigma}}\left(4-\sum_{\sigma}q^{\sigma}\right)^{g-1} \qquad (A.9)$$

となる。

$$\tilde{B} \equiv 4 - \sum_{\sigma} q^{\sigma} \tag{A.10}$$

$$\tilde{m} \equiv \frac{\sum_{\sigma} q^{\sigma}}{(3 - \sum_{\sigma} q^{\sigma})\tilde{B}} \tag{A.11}$$

とおくと

$$\overline{K}(g) = \tilde{m}\tilde{B}^{g} \tag{A.12}$$

とかける。すなわち、$\overline{K}$は$\tilde{B}$を底として指数関数的に増大する。その時の係数が$\tilde{m}$である。 子数分布が二項分布または Poisson 分布の場合、式（A.6）を用いると、

$$\sum_{\sigma} q^{\sigma} = \frac{1}{N^{\varphi}} + \frac{1}{N^{\sigma}} \tag{A.13}$$

となる。たとえば $(N^{\sigma},\ N^{\varphi})$=（1, 2）なら$(\tilde{B}, \tilde{m})$= $(\frac{5}{2}, \frac{2}{5})$ であり、（2, 4）なら $(\tilde{B}, \tilde{m}) = (\frac{13}{4}, \frac{4}{45})$ となる。以下、$(N^{\sigma},\ N^{\varphi})$=（3, 6）なら$(\tilde{B}, \tilde{m}) = (\frac{7}{2}, \frac{2}{35})$, $(N^{\sigma},\ N^{\varphi})$ =（4, 8）なら$(\tilde{B}, \tilde{m}) = (\frac{29}{8}, \frac{8}{203})$と続く。一般的な子数分布の場合，式（A.5）を扱う必要がある。子数分布が解析的に求められない場合や数値計算の場合は$\langle c \rangle^{\varphi}$,$\langle c^2 \rangle^{\varphi}$,$\langle c \rangle^{\sigma}$, $\langle c^2 \rangle^{\sigma}$を測定して代入すればよい。なお、$\langle c^2 \rangle^{\sigma}$が$N^{\sigma}$に比例しない限り、$N^{\sigma}$が大きい時$q^{\sigma}$はゼロに漸近し、$\bar{k}(g) \propto 4^{g}$となることがわかる。