平成２５年度　修士論文

家系図ネットワークにおける継承過程と構造解析

大阪府立大学大学院　工学研究科
電子・数物系専攻　数理工学分野
非線形力学研究グループ

学籍番号　2120103003
生田　成望

2014年2月

目 次

第 1 章　はじめに **3**

第 2 章　Derrida らの先行研究 **6**

2.1 ウェイト 6

2.2 Derrida らの用いた中立モデル 7

2.3 非先祖率 13

第 3 章　モデルによる解析 **14**

3.1 関係性 14

3.2 遺産 22

3.2.1 Derrida らが用いたモデルによる遺産の解析 22

3.2.2 異なる子数分布のモデルの解析 29

第 4 章　実データの解析 **37**

4.1 実データの取得と不完全性 37

4.2 世代の定義づけ 41

4.3 有効なデータの範囲 43

4.4 子数分布 49

4.5 親子間の年齢差 53

4.6 関係性 59

4.7 遺産 63

第 5 章　まとめと今後の課題 **68**

5.1 まとめ 68

5.2 今後の課題 69

謝辞 **70**

	定義 ・ 説明
家系図	生物個体を点で、親子関係にある個体を線で結んだ図
森	生物集団全体の親子関係を結んだネットワーク
主個体	森の中で注目する個体
木	主個体から直接の子孫（あるいは 祖先）を取り出したネットワーク[1]
パス	個体同士を結ぶ線
γ	先祖個体
α	子孫個体
G	主個体から過去方向に辿る世代数
g	主個体から未来方向に辿る世代数
$N(G)$	世代 G の個体数
$\mathrm{I}(G)$	世代 G に属する個体の集合
m	各世代の平均子数
$w(\gamma,\alpha)$	子孫 α に対する祖先 γ のウェイト
$r(\gamma,\alpha)$	祖先 γ に対する子孫 α の関係性
$h(\gamma,\alpha)$	祖先 γ に対する子孫 α の遺産
$\langle q\rangle_c$	条件 c での量 q の平均
$\mathrm{P}(q\|d)$	条件 d での量 q の分布関数
g_M	最短世代
g_Y	年代世代

表 1: 本論文中に現れる用語の説明。

[1]「木」と表記されているが、ループ構造をもっていることに注意。

第1章　はじめに

家系図とは親子関係にある生物個体を線で結んだ図である。一口に家系図と言っても、ある個体の子孫方向のみに注目した図や戦国武将に挙げられるような父子のつながりのみに注目した図など、研究の目的に応じて見方を変えられるほど家系図は多様かつ複雑に構成されている。そのため、本論文で用いる家系図に関わる用語や記号の定義を表1に記す。

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。つまり主個体のG世代前の祖先数をN_Gとすると、

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

となるように思われる。しかし、そのまま増加し続ければ過去のある時代の総個体数を上回ってしまう。例えばヒトの場合、1世代を30年と仮定すると30世代前の先祖の数は式(1.1)から$2^{30} \approx 10$億人と見積もることが出来るが、900年前の世界人口の推計値を大きく上回る[1]。

図1.1は有性生物集団の模式的な家系図の一例である。実際の家系図では、図1.1左の赤で示した個体のように複数の個体の親である、といった役割の重複が存在していると考えられる。この重複した役割を持つ個体の数が世代を遡るにつれて増大し、主個体の祖先数の増加が抑制されるが、それにともない木の構造も複雑になっていくと考えられる。この構造に関する研究がいくつか報告されている。Derridaらは祖先方向に辿る木の構造を特徴づける量をいくつか定義し、中立的なモデルを用いてそれらを解析した[2, 3]。また、水口、西村や堀内は実データの家系図の解析を行った[4, 5, 6, 7, 8, 9, 10, 11]。

また、我々が森と呼ぶ複雑に入り組んだ親子関係を結ぶネットワークは、Derridaらが行ったように家系図を先祖方向にのみ辿る、あるいは本研究で行ったように子孫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）の一種とみなすことが出来る。DAGの一例としては、学術

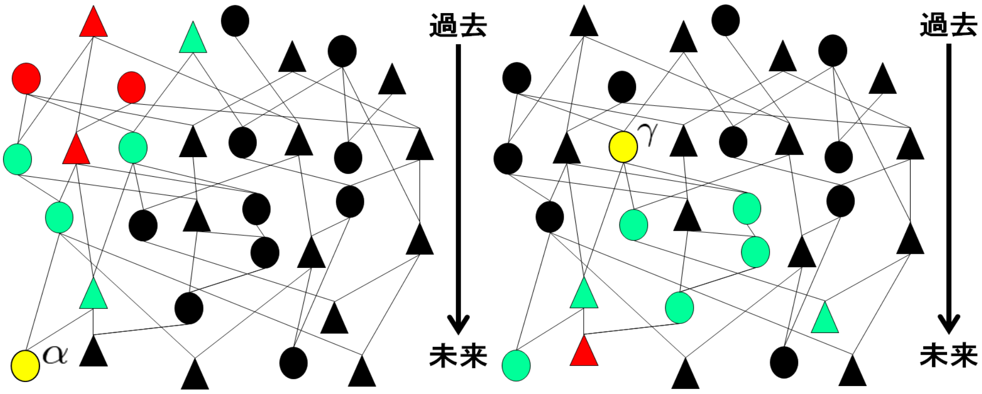


図 1.1: 森と祖先方向に辿った木（左図)、子孫方向に辿った木（右図)。黄色で示した個体 α が主個体、緑と赤で示した個体が α の先祖個体 γ を表す（左図)。黄色で示した個体 γ が主個体、緑と赤で示した個体が γ の子孫個体 α を表す（右図)。丸と三角はそれぞれオスとメスを表し、赤で示した個体は木の中で役割の重複が起きていることを表す。

論文の参照ネットワークが挙げられる。研究はネットワークの構造の特徴付けに限らず、ネットワーク構造の中に存在する特徴的なノードを抽出しようとする動きもある。参照ネットワークにおいて、論文の価値は一様ではなく、参照ネットワークのノード―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。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 [12, 13]。

Derrida らが行っていた研究は注目する子孫個体を選び、その個体から祖先方向に木を定めた場合の家系図の構造に関するものであった。本研究においても家系図の構造を解析するが、家系図の辿り方、すなわち木を伸ばす方向を子孫方向に定めて木を構築し、そのネットワーク構造を解析した。祖先数を例に挙げて役割の重複について述べたが、この役割の重複は図 1.1 の右図の赤色で示した個体のように、子孫方向においても同様に考えられる。ただし、祖先方向に辿る場合のパスの数―すなわち親の数―は常に 2 であるが、子孫方向に辿る場合のパスの数は子数分布に従って確率的に決まることに注意しなければならない。本研究では、Derrida らの用いたウェイトという祖先方向にパスを辿る際に用いる量を、子孫方向に辿る場合に合わ

せて解釈した 2 つの量を定義し、これらの量を用いて Derrida らの用いた中立モデルの解析を行った。

また、Derrida らの用いたモデルは幾つかの仮定に基づき構成されたモデルであり、水口らの競走馬に対する解析によって、実データ一少なくとも競走馬一ではその仮定は破れているとの報告がある。本研究でも、収集した競走馬のデータから家系図を構築し、子孫方向に辿るという方法で木を構築し、競走馬の子孫方向に辿る木と、森全体に関わる性質を解析し中立モデルとの比較を行った。

競走馬のような、長期間に渡って生物集団の発展を解析できる情報の取得は難しい。しかし、他にも例は報告されている。例えば、松尾は近代以前の古文書を紐解き、家族構成と富との間の関係を解析した [14]。

本論文の構成は以下の通りである。第 2 章で Derrida らの先行研究を紹介する。特に、本研究に大きくかかわるウェイトと中立モデルに重点をあてる。第 3 章ではウェイトに対応する二つの量を用い、Derrida らの中立モデルを含む幾つかのモデルを対象とし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 章で本研究のまとめと今後の課題について述べる。

第2章　Derridaらの先行研究

Derridaらは家系図を主個体から祖先方向に辿ることで木を抽出し、木の構造を統計的に調べることで、森すなわち家系図全体の構造を特徴づけた。そのために、彼らはウェイトという量を定義し、中立的な生物集団モデルを用いて木の解析を行った。本章ではこれら、ウェイト・中立的な生物集団モデル・非先祖率の説明と解析結果を紹介する。

2.1　ウェイト

彼らはある注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測る量としてウェイトを定義した。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再帰的に定義される。

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

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

具体的にウェイトを計算するには、まず、主個体に1を与え、その他の全ての個体に0を与える。次に、主個体から過去方向にパスを辿り、自身の持つ量の半分をそれぞれ割り当てる。つまり主個体の両親のウェイトは1/2、祖父母は1/4、...、G世代前の祖先の持つウェイトの値は$1/2^G$、... となるが、役割の重複が発生した場合には子孫らの持つウェイトの半分の総和になる（図2.1参照)。考えている生物集団の中で主個体の祖先を全て辿り終えたら、その時の各個体の値がその個体のウェイトである。また、ウェイトは正またはゼロの実数であり、個体γが主個体αの直接の祖先であれば正のウェイトを持つが、直接の祖先でなければウェイトの値は0となる。

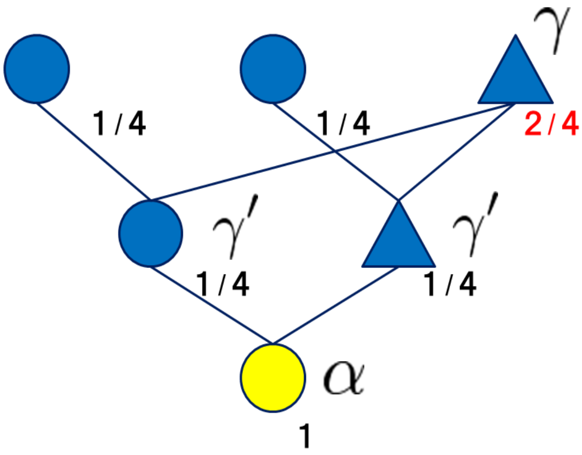


図 2.1: ウェイトの計算例。個体の右下の値はそれぞれの持つウェイトを表す。先祖個体 γ の主個体 α に対するウェイトは $w(\gamma, \alpha) = 2/4$ となる。

子は両親から半分ずつ遺伝子を受け取るため、ウェイトは主個体 α と先祖個体 γ との間にある血縁関係の近さを表現する量と考えることが出来る。と同時に、主個体をスタート地点として、両親のいずれか一方をランダムに選択しながら先祖方向にのみパスを辿る時、主個体からある先祖個体への到達確率であると考えることが出来る。

2.2 Derrida らの用いた中立モデル

実際の生物集団は第 1 章の図 1.1 で示した通り、出生年が個体毎に異なり世代の定義が難しい。彼らは世代の重複が起こらないガルトン＝ワトソン型の単純な生物集団モデルに対して木の構造解析を行った。彼らが構造解析に用いた中立的な生物集団のモデルは以下の 4 つの条件を満たす。

1. 世代間の重複はない
2. オスとメスが同数
3. ランダムに対をなし子を作る
4. 対あたりの子数は雌雄ともに平均 m のポアソン分布に従う

主個体 α の属する世代を $G=0$ と定義し、$G=1$ は $G=0$ の親の世代、$G=2$ は $G=0$ の祖父母の世代、...、を表す（図 2.2 参照）。また平均 m のポアソン分布

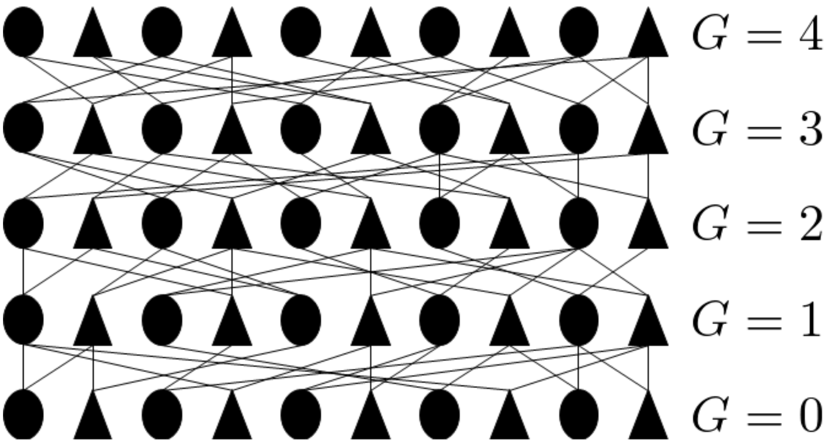


図 2.2: モデルによって作られた仮想的な生物集団の家系図の例。●と▲がそれぞれオスとメスを表している。

$P(k)$ は

$$P(k)=\frac{m^k \mathrm{e}^{-m}}{k!}\ (k=0,\,1,\,2,\ldots)\ , \tag{2.3}$$

で表される。実例として、1350 年から 1986 年のイギリスにおけるヒトの夫婦が k 人の息子を持つ時の確率分布が 6 人までは平均 1.15 のポアソン分布でよくあらわされることが知られている（図 2.3 参照)。中立モデルにおける条件の一つ、“ ランダムに対をなし子を作る ”はいわば一夫一妻制であり、兄弟姉妹は両親が一致している。しかし、この条件を変更し、図 2.2 のように配偶する相手を固定せず、子の方から一つ上の世代の個体からランダムに親を選択するというアルゴリズムを用いてもこれから示す結果に影響はない。

彼らは 2.2 節で紹介したモデルを用い、本節で紹介したウェイトと、各世代における α の先祖ではない個体の割合を指す非先祖率という量を統計的に解析することで、家系図の構造を特徴づけた [2, 3]。非先祖率は次節で紹介する。

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はある形に収束し、その分布はウェイトの値が微小な範囲においてべき則を示すことが示された (図 2.4 参照)。図 2.4 の P (w) は規格化されたウェイト w の分布関数を f (w) とすると、P (w) = f (w) /w で与えられる。これは実際の確率分布をウェイトの値で割った

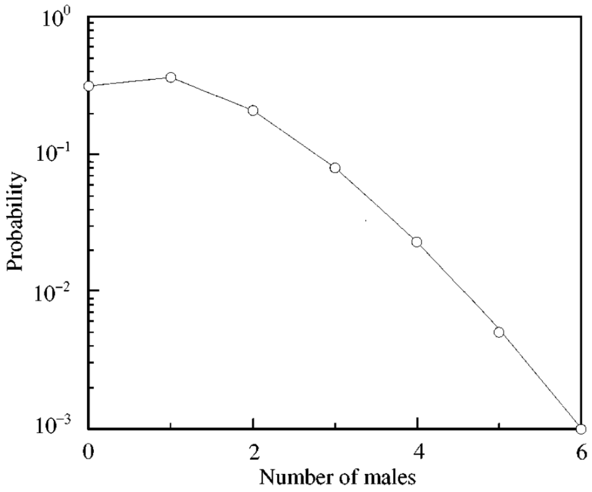


図 2.3: 1350 年から 1986 年の期間におけるイギリスの夫婦が k 人の息子を持つ時の確率分布。丸印で表されている。横軸は息子の数、縦軸は確率分布を表す。縦軸は log スケールである。参考文献 [3] より転載。

数である。この分布におけるべきの指数は 0.3 である。なお、本論文では Derrida らによって得られたウェイトの分布を『D 分布』と呼ぶ。

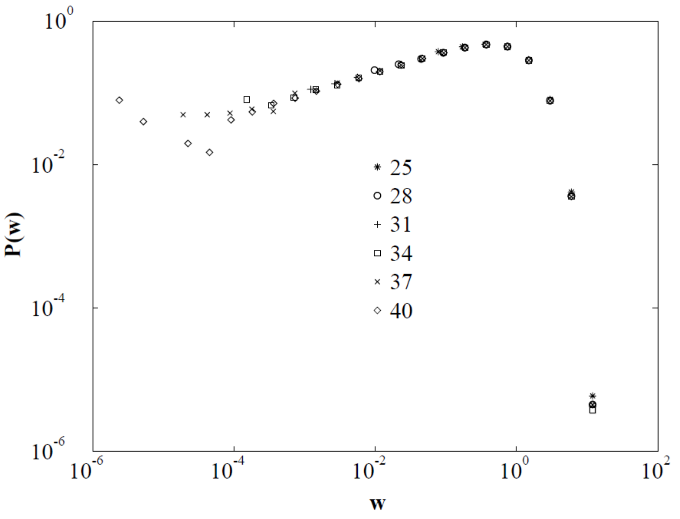


図 2.4: 各世代の個体数が 2^{20}（平均子数 $m = 2$）のモデルにおけるウェイトの分布。横軸は規格化されたウェイトの値を表す。縦軸は w の分布関数を f (w) として P (w) = f (w) /w で与えられる。異なる記号は α から遡った世代別の分布を表す。

また、彼らは森の中の祖先方向に辿る木同士の近さを特徴づけるために以下の量を用いた。

$$X^{(\alpha)}(G) \equiv \sum_{\gamma \in \mathrm{I}(G)} [w(\gamma, \alpha)]^2, \tag{2.4}$$

$$Y^{(\alpha_i, \alpha_j)}(G) \equiv \sum_{\gamma \in \mathrm{I}(G)} w(\gamma, \alpha_i)\, w(\gamma, \alpha_j)\ . \tag{2.5}$$

$\mathrm{I}(G)$ は主個体がいる世代から G 世代前の世代に属する個体の集合を表す。$X^{(\alpha)}(G)$ は、世代 G の $\gamma_1, \gamma_2, \ldots, \gamma_{N(G)}$ が持つ、ある α を主個体にしたときのウェイトの値を変数とした $N(G)$ 次元ベクトルである。$Y^{(\alpha_i, \alpha_j)}(G)$ は主個体が α_i、α_j のときの 2 つのベクトルの内積である。この 2 つの量を用いて α_i、α_j から祖先方向に辿る二本の木の重複度合いを求める量 $q^{(\alpha_i, \alpha_j)}(G)$ を定義した。

$$q^{(\alpha_i, \alpha_j)}(G) \equiv \frac{Y^{(\alpha_i, \alpha_j)}(G)}{[X^{(\alpha_i)}(G)\, X^{(\alpha_j)}(G)]^{1/2}}\ . \tag{2.6}$$

$q^{(\alpha_i, \alpha_j)}(G)$ は 2 つのベクトルのなす角の余弦であり、その値が 1 に近づくほど、α_i を主個体とした時の γ らが持つウェイトと、α_j を主個体とした時の γ らが持つウェイトに差が無いことを示す。中立モデルに対して $q^{(\alpha_i, \alpha_j)}(G)$ の α_i、α_j に対する平均値 $\mathrm{q}(G)$ の G 依存性は図 2.5 のようになった。すなわち、十分世代を経るとウェイトは子孫個体に依らない量であることが分かった。

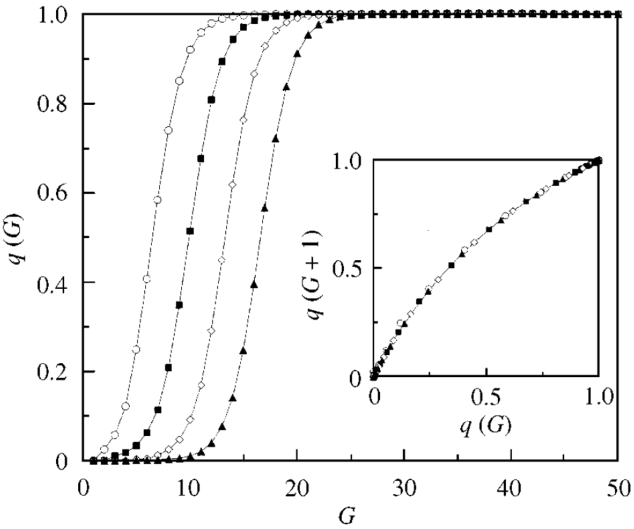


図 2.5: 森のサイズを変更した時の $\mathrm{q}(G)$ の推移。白丸が各世代の個体数が 100、黒四角が 1000、白四角が 10000、黒三角が 100000 である。

2.3 非先祖率

解析を行ったもう一つの量である非先祖率 $S(G)$ とは、α の属する世代を基準とし、α の属する世代から G 世代前の個体の中で α の非先祖が占める割合、つまり G 世代前における $w(\gamma,\alpha)=0$ の個体の占める割合である。α の属する世代においては α を除いた個体全ての割合であり $S(0)=1-1/N(0)$ となる。また 1 世代前は両親のみであるため、$S(1)=1-2/N(1)$ となる。$G>1$ の世代においては役割の重複が起きるため、2 の指数で先祖が増加し続けることはなく単純に $S(G)=1-2^G/N(G)$ と記述が出来ない量である。この量は Derrida らの用いた中立モデルにおいて

$$S(G+1) = \exp[-m+mS(G)], \tag{2.7}$$

$$S(0) = 1-\frac{1}{N(0)}, \tag{2.8}$$

に従うことを理論的に示し数値的に確かめた（図 2.6）。ここで、$N(0)$ は主個体の属する世代にいる個体数を指す。ウェイトの分布と同様に十分世代を経ることで収束し、人口増減の無い $m=2$ の時には約 0.2 に収束する。

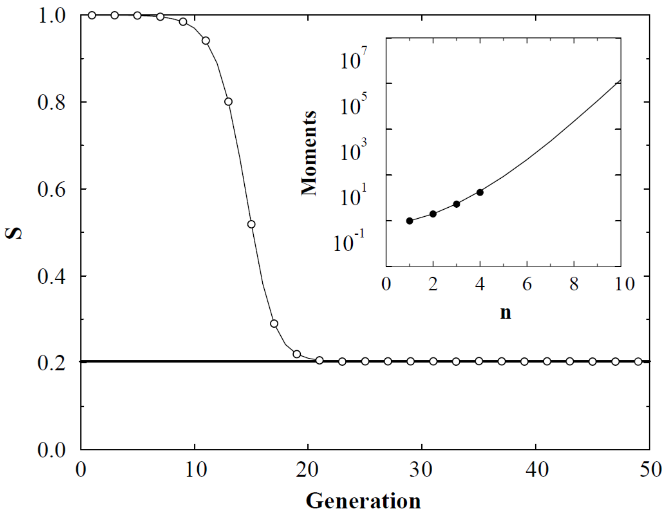


図 2.6: $N(G)=2^{15}$（$m=2$）のときの非先祖率の世代依存性。横軸が遡った世代数、縦軸が非先祖率を表す。参考文献 [2] より転載。

第3章　モデルによる解析

前章で紹介したDerridaらによる先行研究では、家系図の構造を特徴づけるために、主個体から親子関係を祖先方向に辿ることによって構成される木に着目した。ウェイトは木に属する個体のみ正の値を持つ量であり、逆にウェイトがゼロの個体の割合が非先祖率である。本研究では、子孫方向に辿ることによって構成される木に着目する。本章では、Derridaらが用いた中立的な生物集団モデルを対象として解析した。子孫方向に辿る場合にウェイトに対応する量は二種類考えられる。3.1節では血縁関係の近さを表現する量である関係性についての解析結果を、3.2節では到達確率を表現する量である遺産についての解析結果をそれぞれ報告する。

3.1　関係性

家系図の中で注目する先祖個体を主個体と呼びγで表し、γのある子孫個体をαと表す。このとき、子孫個体αの主個体γに対する関係性$r(\gamma,\alpha)$を以下のように再帰的に定義する。

$$r(\gamma,\gamma) \equiv 1\,, \tag{3.1}$$

$$r(\gamma,\alpha) \equiv \sum_{\alpha\text{の両親個体}\alpha'} \frac{r(\gamma,\alpha')}{2}\,. \tag{3.2}$$

具体的に関係性を計算するには、まずγ自身の値を1とし、γ以外の個体には0を与える。次に、γから家系図を子孫方向にのみ辿り、γの子孫らにそれぞれの親個体が持つ量の半分を与えていく。これは子が両親から半分ずつ遺伝子を受け継いでいることと対応しており、ウェイトの持つ意味の一つである血縁関係の近さを表している。子の持つ量は両親から受け継ぐ量の和になる（図3.1参照）。γの全ての子孫個体を辿り終えた時の値が関係性となる。αがγの子孫でない場合は$r(\gamma,\alpha)=0$であり、また、その値が大きければ大きいほど主個体と遺伝子的に近い関係であるこ

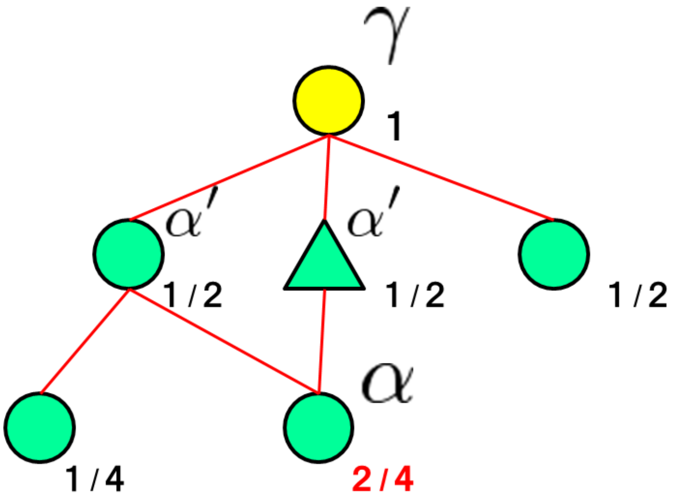


図 3.1: 森の中で γ から子孫方向に辿った木。個体の右下に書かれた数字はその個体の持つ関係性の値を表す。

とを表すため、ここで定義した量を関係性と名付けた。両親は必ず二個体だが子数は確率的に決まり、選択した γ によっては血が絶えることに注意が必要である。そのため、関係性はウェイトと異なり、世代間で非保存量である。各世代での γ の子孫の平均子数が 2 の場合を除き各世代における関係性の総量は変化する。

我々は、2.2 節で述べた中立生物集団モデルに対して、この関係性の分布とその世代依存性を調べた。具体的には各世代で雌雄同数を用意し、各個体が一つ過去の世代から両親をランダムに選び取るという方法を用いて作成した。この場合、2.2 節の条件と異なり、配偶関係は多夫多妻となる。しかし、平均値や分布などの統計量に関する解析結果については大きく影響を与えないと考えられる。本研究では、子数分布がポアソン分布を示す生物集団においては全てこの方法を用いて解析を行った。

関係性は各世代において総和が非保存な量であることから、まず主個体毎に異なるであろう各世代における関係性の平均値に注目することにした。関係性 $r(\gamma, \alpha)$ は、γ と α によって決まる量であり、その統計量を考えるときには、どういう母集団を取るかに注意しなければならない。最初に注目したのは、γ を固定し α の母集

団として g 世代に属する個体の集合 $\mathrm{I}(g)$ を取った場合の関係性の平均値：

$$\langle r(\gamma,\alpha)\rangle_{\{\alpha\}=\mathrm{I}(g)} \equiv \frac{1}{N}\sum_{\alpha\in\mathrm{I}(g)} r(\gamma,\alpha) \tag{3.3}$$

である。記号 $\langle\cdot\rangle$ は平均値を表し、添字 $\{\alpha\}=\mathrm{I}(g)$ は、α を母集団として $\mathrm{I}(g)$ をとることを示している。また、本モデルでは世代 g の個体数 $N(g)=N$（一定）である。子は親の量の半分かつ個体数は各世代同数であるから、ある世代の γ の子孫らの平均子数が生物集団の平均子数 2 を越えると、次世代の式 (3.3) の値は増加し下回ると減少する。作成した森の中で最も過去世代から主個体を選択し、家系図を子孫方向に辿り全個体の関係性を計算し平均値を求めた。典型的な例として $\gamma_1,\gamma_2,\gamma_3$ に対する平均値 (3.3) の世代依存性を図 3.2 に示す。図 3.2 にある $\langle r(\gamma,\alpha)\rangle_{\{\alpha\},\{\gamma\}}$

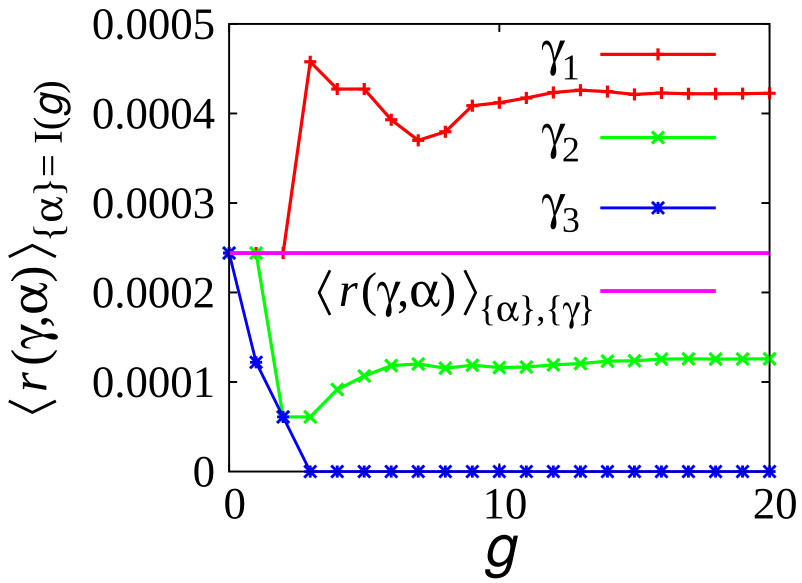


図 3.2: 各世代における関係性の平均値の典型例と平均。横軸が世代、縦軸が関係性の平均値。γ_1、γ_2、γ_3 は $\mathrm{I}(0)$ に属する個体であり、$\langle r(\gamma,\alpha)\rangle_{\{\alpha\},\{\gamma\}}$ は $\mathrm{I}(0)$ の個体らから求めた式 (3.3) の平均値を指す。$N=4096$。

は $\mathrm{I}(0)$ の母集団全体の平均値を指しており、

$$\langle r(\gamma,\alpha)\rangle_{\{\alpha\},\{\gamma\}} \equiv \frac{1}{N}\sum_{\gamma\in\mathrm{I}(0)} \langle r(\gamma,\alpha)\rangle_{\{\alpha\}=\mathrm{I}(g)} \tag{3.4}$$

で表される。各個体で見ると子数はポアソン分布に従うため確率的に決まるが、生物集団全体で見ると子数は m とみなせる。そのため、世代 g に受け継がれる関係

性の総量は$m^g/2^g$となる。また、世代gの個体数は$g=0$の個体数$N(0)$を用いて$N(0)\times(m/2)^g$と書けるから、本論文で限定している$m=2$の場合に限らず母集団全体の平均値は常に$1/N$となる量である。図 3.2 のγ_1とγ_2の結果から、母集団から選択するγによって受け継がれる関係性の量が大きく異なることが分かる。図 3.2 のγ_3は$g=3$で子孫が絶えてしまう個体で、血が絶えてしまう個体は$\mathrm{I}(0)$の中で二割ほど存在する。これは前述の非先祖率の解析にて示された結果である、$m=2$においてGが十分大きい$\mathrm{I}(G)$の個体らはαと血のつながりのない個体がおよそ二割存在することと無矛盾である。図 3.2 の$g>10$の範囲でγ_1とγ_2共に値が一定になっているように見えるが、これは血が絶えないγの子孫は、世代を辿り続けると$\mathrm{I}(g)$の個体の大部分、いずれは全てを占めるためである。関係性の定義から、両親が共に関係性の値を持つと子の持つ関係性の量は両親の中間の値になり、配偶がランダムであることから、世代が経つにつれて集団全体で均質化され同世代での個体差は無くなっていく。これは図 3.3 や図 3.4 から見て取れる。図 3.3 はγ_1とγ_2から受け継がれる$\mathrm{I}(20)$と$\mathrm{I}(22)$の個体が持つ関係性の分布である。図 3.3 のγ_1とγ_2は図 3.2 で選択したものと同一の個体である。これを見るとγ_1とγ_2はともに正規分布でよくフィッテングが出来ることが分かり、世代が経つことで分散が小さくなることが分かる。更に世代が経過することでδ関数に漸近する。図 3.4 はγ_1とγ_2から受け継がれる関係性の分布の分散の世代依存性である。始めの数世代においては図 3.2 同様に選択した個体による違いが見られるが、gが大きくなるとどれも同じ傾きを示す。これも$\mathrm{I}(g)$の個体の大部分、いずれは全てを占めるためと考えられる。$\mathrm{var}(0)\times 2^{-g}$は、母集団を$\mathrm{I}(0)$ととり求められる$\gamma$毎の関係性の分散の平均値によくフィッティングできる関数である。$\mathrm{var}(0)$は$g=0$における関係性の分布の分散を表しており、$g=0$において主個体関係性は 0 であるため、この値は式 (3.5) と書ける。

$$\mathrm{var}(0)=\frac{1}{N}\sum_{\alpha\in\mathrm{I}(0)}\left(h(\gamma,\alpha)-\langle\, r(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\},\{\gamma\}}\right)^2=\frac{(N-1)}{N^2}\,. \tag{3.5}$$

分散の減衰は指数2^{-1}の指数関数で上手くフィッティングできるが、この傾向が平均子数mに依存するかどうかを調べるためにmを変えて分散の平均値を計測した (図 3.5 参照)。その結果、どの平均子数の森においても分散の平均値に違いが見られなかったため、$\langle\, r(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\},\{\gamma\}}$と同様にこのモデルにおいては分散の平均値は$m$に依存しない関数で表すことができると結論付けた。

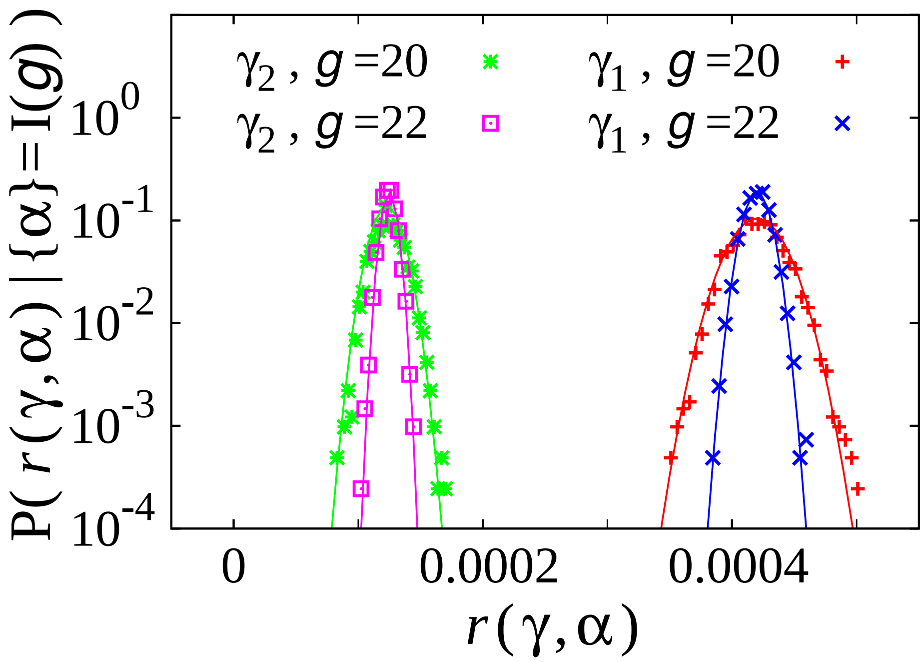


図 3.3: $N = 4096$ の森において、γ_1 と γ_2 から受け継がれる I (20) と I (22) の個体が持つ関係性の分布。横軸は I (20) または I (22) に属する α が持つ関係性であり、縦軸はその割合を指す。プロットはそれぞれの計測値で、実線はそれぞれを正規分布でフィッティングしたものを指す。

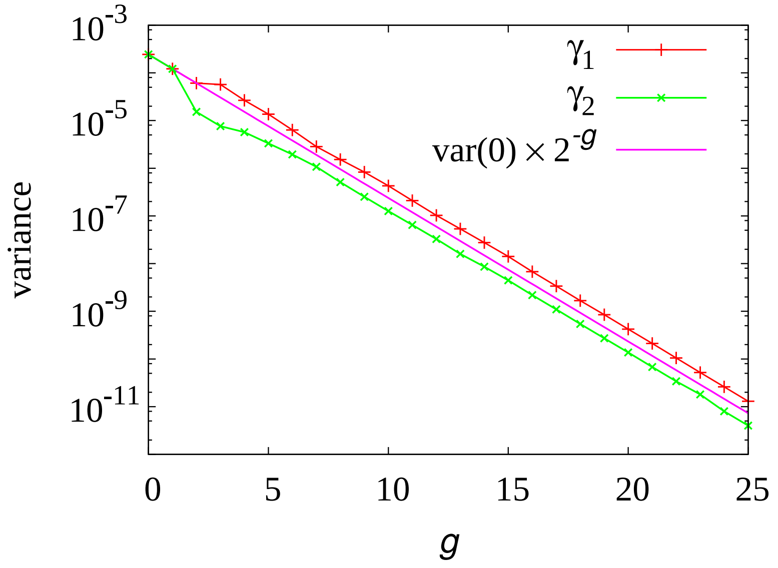


図 3.4: $N = 4096$ の森において、γ_1 と γ_2 から受け継がれる関係性の各世代の分散。横軸は世代を指し、縦軸は分散を表す。$\mathrm{var}\,(0)$ は 0 世代の $r\,(\gamma, \alpha)$ の分散である。

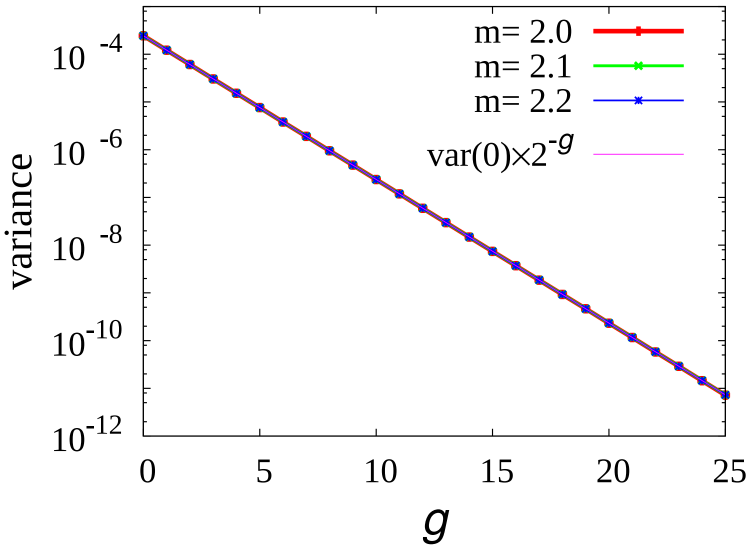


図 3.5: $N\,(0) = 4096$ の森において、森の平均子数 m を変えたときの γ を母集団とした関係性の分散の平均値。横軸は世代を指し、縦軸は分散を表す。$\mathrm{var}\,(0)$ は 0 世代の $r\,(\gamma, \alpha)$ の分散である。

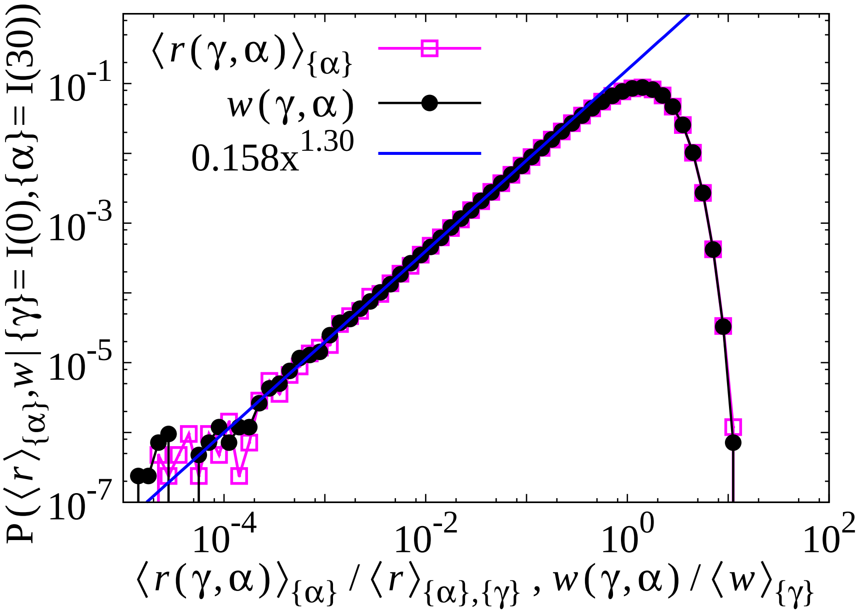


図 3.6: $N=4096$ の森における I(0) の個体 γ が、I(30) 中の子孫 α へ与える関係性の平均値 $\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ の分布（紫四角)。γ の母集団として I(0) を取りかつ、乱数のシードを変えて作った 1000 通りの森を母集団として平均を取ったもの。比較のために、I(30) に属するある個体 α から求められる I(0) の個体が持つウェイトの分布を黒丸で描いている。横軸は規格化された関係性の平均値とウェイトを表し、縦軸はその分布を表す。直線は $\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ 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。

次に、$\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ の γ 依存性に着目した。すなわち、十分大きい g において、ある γ から受け継がれる関係性は I(g) に属する個体 $\alpha_1,\alpha_2,\ldots,\alpha_N$ において差は無いが、$\gamma_1,\gamma_2,\ldots,\gamma_N$ によって I(g) に受け継がれる関係性には大きな差がある。$\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ の γ の依存性を知るため、I(0) の個体 γ らがそれぞれ受け継がせる $\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}$ の γ を母集団とした分布 $\mathrm{P}\left(\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}|\{\gamma\}=\mathrm{I}(0)\right)$ を求めた。その結果を図 3.6 に示す。これまでの図はある乱数によって生成された一つの森に対して得られたものであるが、分布 $\mathrm{P}\left(\langle r(\gamma,\alpha)\rangle_{\{\alpha\}}|\{\gamma\}=\mathrm{I}(0)\right)$ は γ に大きく依存するとともに作成した森によっても取りうる値にはばらつきが見られるため、図 3.6 では乱数のシードを変えることによって作った 1000 通りの森についても平均を取った。これに対して、図 3.2、3.3、3.4、3.5 は単一の森に対して得られた結果であるが、これらの図で示

された特徴は、森の選び方に依らない。また先行研究で示されたウェイトの分布との比較を行うために $\langle r(\gamma, \alpha)\rangle_{\{\alpha\}=\mathrm{I}(30)}$ を $\langle r(\gamma, \alpha)\rangle_{\{\alpha\},\{\gamma\}}$ で規格化した。この分布はサイズに依らず N を変えても分布の形は大きく変わらない。図 3.6 のウェイトの分布を求めるに当たり、図 2.5 から十分世代を経る時ウェイトの分布は α に依らないという結果が得られていたため、関係性の平均値の分布を求めた各森で $\mathrm{I}(30)$ の中から一つの α を選択し、$\mathrm{I}(0)$ を母集団としたウェイトの分布を求めて平均を取った。図 3.6 では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。先行研究においてべきの指数は 1.2991138 . . . と述べられており、図 3.6 でフィッテングした指数 1.30 は、先行研究の結果とよく一致する。 2.1 節の図 2.4 で紹介した分布と図 3.6 で示したウェイトの分布とでは形が異なるように見えるが、図 2.4 の縦軸 $\mathrm{P(w)}$ は、分布関数そのものではなく、ウェイトの逆数をかけた $\mathrm{P(w)/w}$ だからである。

関係性の分布とウェイトの分布は非常によく似ているが、これは関係性の定義によるものである。ウェイトの値と関係性の値はともに γ と α を結ぶ親子関係のパスの数に 2^{-g} をかけた値であり、選択した γ と α が同一であればウェイトと関係性は全く同じ大きさになる。そのため、本研究によって得られた γ から受け継ぐ関係性は、十分大きい世代 g に属する個体間では差が無いこと、また母集団を $\mathrm{I}(0)$ にとった時の十分大きい世代 g における $\langle r(\gamma, \alpha)\rangle_{\{\alpha\}=\mathrm{I}(30)}$ の分布が値の微小な範囲でべき則を示すことは、先行研究で知られている α を変えても十分離れた世代のウェイトの分布が変わらないことと、α から十分離れた世代の個体が持つウェイトの分布が値の微小な範囲においてべき則を示すこととそれぞれ対応している。図 3.6 において関係性の平均値の分布とウェイトの分布がわずかにずれているが、これはウェイトの分布を求める際、森毎に一つの α しか選ばなかったことに起因する。同じ森を用いてウェイトを求める際に、α を母集団としたウェイトの分布を求めれば、関係性の平均値の分布とウェイトの分布は一致する。また血が絶えてしまい十分大きな g の世代に関係性を受け継ぐことが出来ない個体が 2 割存在することと、ウェイトの非先祖率 S は対応している。

3.2 遺産

本節では、家系図を祖先個体 γ を起点として、各個体の子の中から等確率でランダムに選び、子孫方向に辿る過程を考える。子孫個体 α への到達確率を本論文では遺産と呼び解析に用いた。主個体 γ の子孫個体 α の持つ遺産は以下の式により再帰的に定義される。

$$h(\gamma,\gamma) \equiv 1, \tag{3.6}$$

$$h(\gamma,\alpha) \equiv \sum_{\alpha\text{の両親個体}\alpha'} \frac{h(\gamma,\alpha')}{N_{\alpha'}}. \tag{3.7}$$

$N_{\alpha'}$ は α の親個体 α' の子数を表す。具体的に遺産を計算するには、前節の関係性と同様に、まず主個体 γ 自身の値を 1 とし、γ 以外の個体には 0 を与える。次に、γ から子へ、自身の持つ遺産を等分した量を継承させる。α' は α の親であるため、$N_{\alpha'} \geq 1$ となる。役割の重複が起きた際に子孫が受け継ぐ遺産は、関係性と同様に両親から受け継がれる遺産の和になる（図 3.7 参照)。γ の全ての子孫個体を辿り終えた時の値が遺産となる。親が持つ量を子へ等分配することがいわゆる財産の相続手続きを想起させるため、ここで定義した量を遺産と名付けた。遺産は自身の持つ量を等分配するため、関係性とは異なり子数が 0 でない限り保存する。しかし、遺産を持つ個体が子をなさないとき遺産は継承されず失われる。そのため遺産も非保存量であることに注意が必要である。

3.2.1 Derrida らが用いたモデルによる遺産の解析

我々は、いくつかの種類の個体集団に対して、この遺産の分布とその世代依存性を調べた。最初に解析の対象としたのは、 2.2 節で述べた中立生物集団モデルである。遺産の継承過程を特徴づけるため、主個体 γ に対する世代毎の遺産の平均 $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ を以下のように定義した。

$$\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)} \equiv \frac{1}{N} \sum_{\alpha \in \mathrm{I}(g)} h(\gamma, \alpha). \tag{3.8}$$

関係性の時と同様に世代 g を変えた時の $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ を図 3.8 に示した。子を持たない個体が持っていた遺産は次世代へ受け継がれないため、世代が経つにつれ

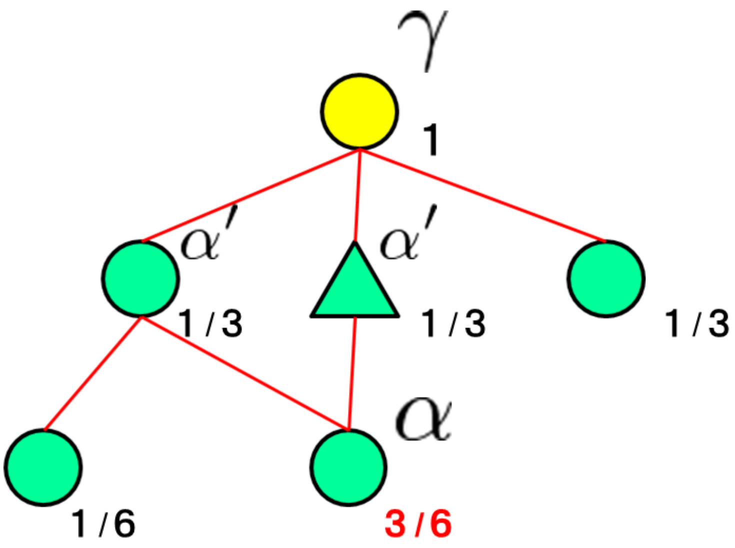


図 3.7: 遺産の計算例。個体の右下の値はそれぞれの持つ遺産を表す。この木において図のように γ と α を取ると $h\,(\gamma,\alpha)=3/6$ となる。

て遺産の総量は単調減少する。生物集団の子数が平均 m のポアソン分布に従う時、生物集団全体で見れば世代毎に e^{-m} の割合で遺産は減少する。このモデルにおいて、生物集団全体で継承される遺産の平均の g 依存性は

$$\langle h\rangle_{\{\alpha\},\{\gamma\}} \equiv \frac{1}{N}\sum_{\gamma\in \mathrm{I}(0)} \langle h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)} = \frac{1}{N}\left(1-\mathrm{e}^{-2}\right)^{g}, \tag{3.9}$$

となる。図 3.8 にある γ_1 と γ_2 の実線が、それぞれ $g=10$ 前後から直線に乗っているが、この直線の指数は式 (3.9) から導かれる指数 $(1-\mathrm{e}^{-2})$ と一致する。また、γ の選び方に依って受け継がれる遺産量がおよそ 100 倍も変化することから、受け継がれる遺産量は先祖個体 γ の選び方に大きく依存することが分かった。

次に図 3.8 で用いた γ_1 に対して、g 世代の個体の持つ遺産の分布を求めた（図 3.9）。図 3.9 は両対数プロットしている。分布の平均値は図 3.8 から世代を経る毎に指数関数的に減衰することが分かるため、各世代における遺産の対数を取った値から求められる平均値 $\langle \log_{10} h\rangle$ を用いて規格化している。図 3.9 中の h は $h\,(\gamma,\alpha)$ を表し

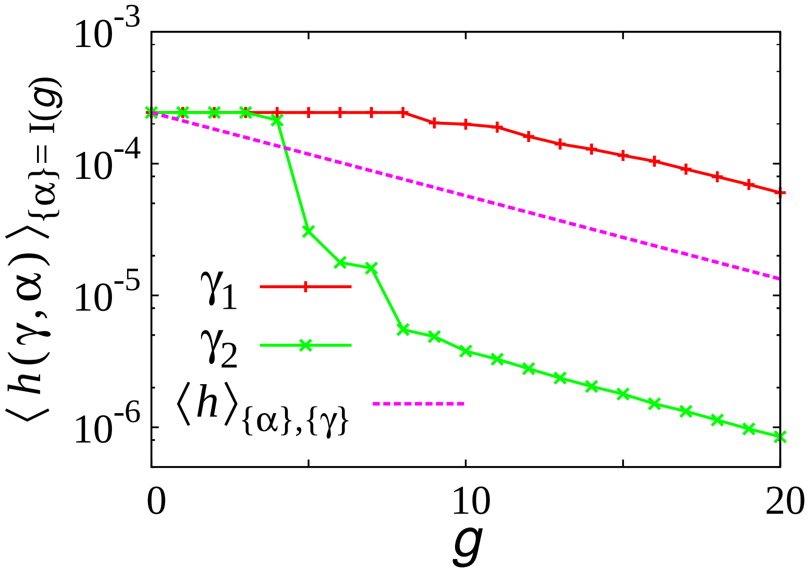


図 3.8: 各世代における遺産の平均値の典型例と平均。横軸が世代、縦軸が遺産の平均値。実線はそれぞれ $\langle\, h\,(\gamma_1\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ 、$\langle\, h\,(\gamma_2\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ を表す。破線は $g=0$ の個体を γ の母集団とした時の $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の平均 $\langle h\rangle_{\{\alpha\},\{\gamma\}}$ である。$N=4096$。

ており、横軸と縦軸で用いられている $\log_{10} h$ は $\log_{10} h = \log_{10} h\,(\gamma,\alpha)$、$\langle\log_{10} h\rangle$ は

$$\langle\log_{10} h\rangle \equiv \frac{1}{N}\sum_{\alpha\in\mathrm{I}(g)} \log_{10} h\,(\gamma,\alpha)\ , \tag{3.10}$$

を表す。図 3.9 より、世代が経つにつれ分布の形が収束していくことが分かる。関係性の分布では関係性の平均値の値が収束している $g=20$ ではっきりと正規分布であることが確認できたが、遺産の分布は $g=20$ ではまだ収束しておらず、関係性の分布に比べ収束に時間がかかることが分かった。

次に収束する分布の形の γ 依存性を調べるため、図 3.8 で用いた γ_1 と γ_2 に加えいくつか γ を選択し、40 世代の個体の持つ遺産の分布を求めた（図 3.10 参照）。図 3.10 は図 3.9 同様に両対数プロットである。選択する γ によって受け継がれる遺産の量が異なるため、図 3.9 と同様に $\langle\log_{10} h\rangle$ を用いて規格化を行ったものについて比較した。これらの γ から求められる分布はよく一致しており、遺産の分布は平均値こそ異なるものの、収束した分布の形は γ に依存しないことが分かった。

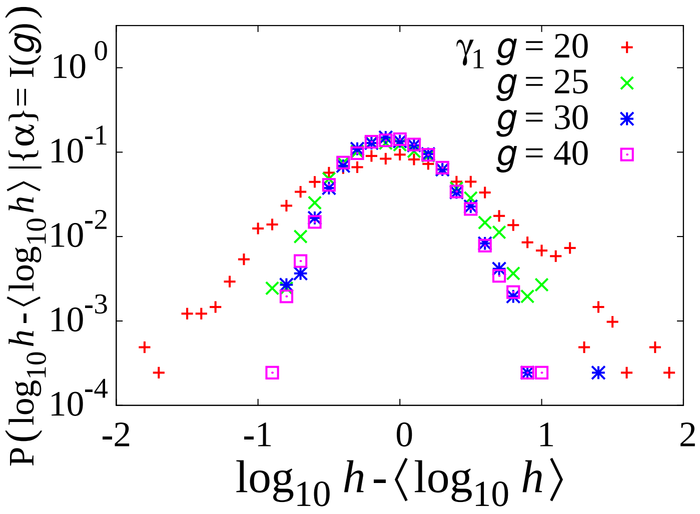


図 3.9: 図 3.8 で示した個体 γ_1 から受け継がれる I (g) の個体が持つ遺産の分布。横軸が遺産、縦軸がその割合を表す。それぞれの世代によって平均値が異なるため、プロットはそれぞれの世代の個体が持つ遺産を規格化した値を示す。

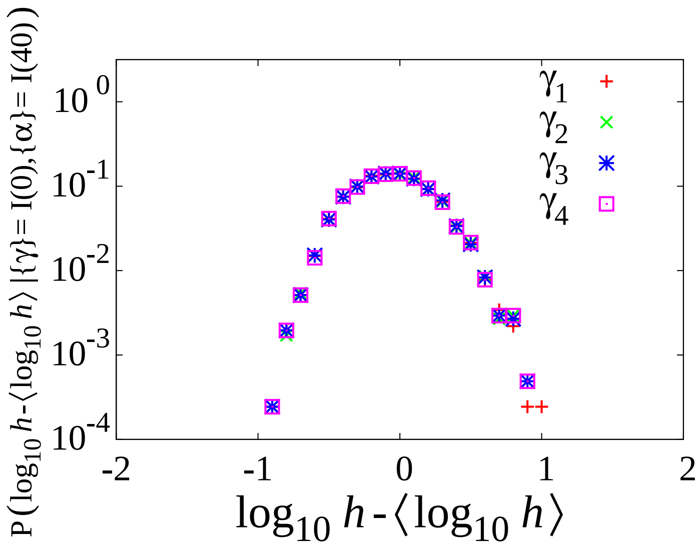


図 3.10: γ_1、γ_2、γ_3、γ_4 から受け継がれる I (40) の個体が持つ遺産の分布。γ_1、γ_2 は図 3.8 で用いた個体、γ_3、γ_4 は I (0) から新たに選択した個体である。横軸が遺産、縦軸がその割合を表す。γ によって平均値が異なるため、プロットはそれぞれの γ から受け継がれる I (40) での遺産の平均値を用い規格化した値を示す。

図 3.9 および図 3.10 で表されている分布は一見すると放物線でフィッティングできるように見える。両対数のグラフが放物線でフィッティングできるとすると、その分布は対数正規分布で表せることになる。しかし、実際にはそうではない。遺産の分布が対数正規分布には従わないことを確かめるため、遺産の対数を取った値を用いて各世代での歪度 S_k を求めた（図 3.11 参照)。

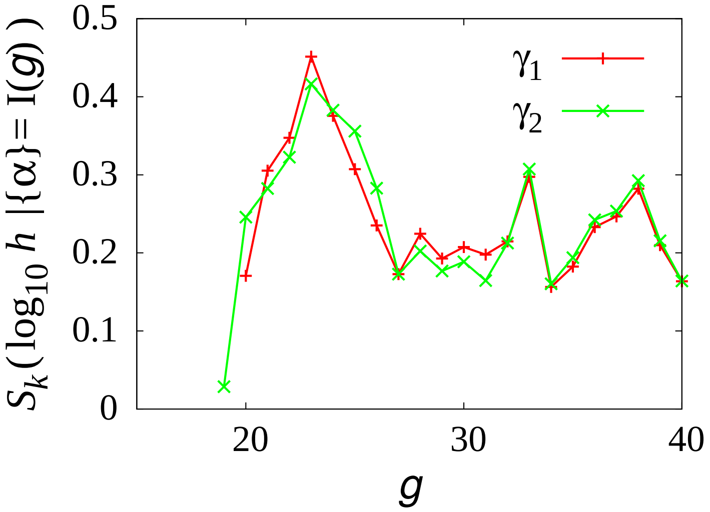


図 3.11: γ_1、γ_2 から受け継がれる各世代での遺産の対数を取った値の歪度。γ_1、γ_2 は図 3.8 で用いた個体である。横軸が世代、縦軸がその世代における歪度を表す。

S_k は以下の式で表される。

$$S_k = \frac{\sum_{\{\alpha\}\in \mathrm{I}(g)} (\log_{10} h - \langle \log_{10} h \rangle)^3}{N^{-0.5}\{\sum_{\{\alpha\}\in \mathrm{I}(g)} (\log_{10} h - \langle \log_{10} h \rangle)^2\}^{3/2}}. \tag{3.11}$$

歪度は本来分布の左右対称性を表す量であり、対数正規分布において歪度は常に 0 でない有限の値を持つ。ただし、今回は遺産の対数を取った値を用いて歪度を求めるため、遺産の分布が対数正規分布で表されるためには、S_k が十分小さくならなければならない。図 3.11 で γ_1 と γ_2 の歪度が $g = 20$ 前後まで存在しないが、遺産の対数を取っており、同世代間の個体が全て遺産を持つまでは歪度の計算が出来ないためである。γ_1 と γ_2 の歪度は始めは異なる値を取っているが、世代を経る毎に γ を変えても同じ歪度を取るように見える。この事実は、十分大きい g において、どの γ から受け継がれる遺産の分布であっても収束した分布の形は一致するという結果を支持する。

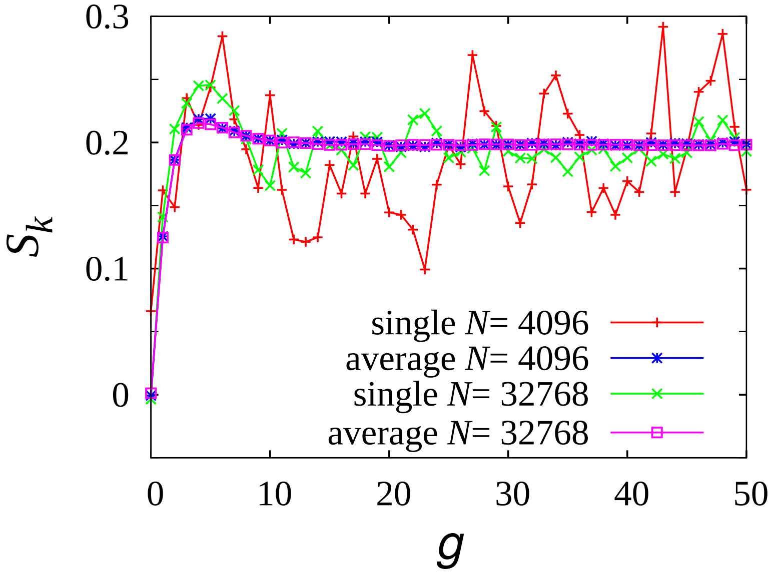


図 3.12: $N = 4096$ と $N = 32768$ における、I(0) の個体に遺産を与えた場合の各世代における S_k の値。横軸は世代を表し、縦軸は各世代における遺産の対数を取った量の歪度を表す。I(0) の個体には $g = 0$ での遺産の分布が $\mu = 10.0$, $\sigma = 10.0$ の対数正規分布になるよう乱数を与えた。single は単一の森の場合の歪度を表し、average は 1000 回森を作成して求めた平均の歪度を表す。

図 3.11 から、どの γ を選択しても、血が絶えない限りは十分大きな g において分布は同じ特徴を示すこと、および遺産の対数を取った量の歪度は 0.2 程度の値を取り続けることが予測される。しかし、歪度が 0 の値を取らない要因は辿る世代数が十分でないため、または N の大きさによる有限サイズの効果が働いているためとも考えられる。そこで、遺産の分布は最終的に対数正規分布になると仮定し、遺産の分布が対数正規分布となるように I(0) に属する個体全てに遺産を与え、その後 $g = 50$ までの遺産の対数の歪度を計算した。結果は図 3.12 である。用いた対数正規分布の μ と σ はそれぞれ $\mu = 10.0$, $\sigma = 10.0$ であり、これらのパラメータに応じて I(0) での遺産の総量や遺産の分布の分散は変化するが、S_k はこれらに依存しない。図 3.12 中の single は単一の森での歪度を示し、average は森を 1000 回作成して平均した歪度を示す。初期条件が対数正規分布になるよう乱数を与えているため、$g = 0$ においてどの結果も $S_k \approx 0$ を示す。分布が対数正規分布であれば S_k は $S_k \approx 0$ であり続けるはずだが、どの結果においても始めの数世代で 0.2 を超える値にまで増加する。

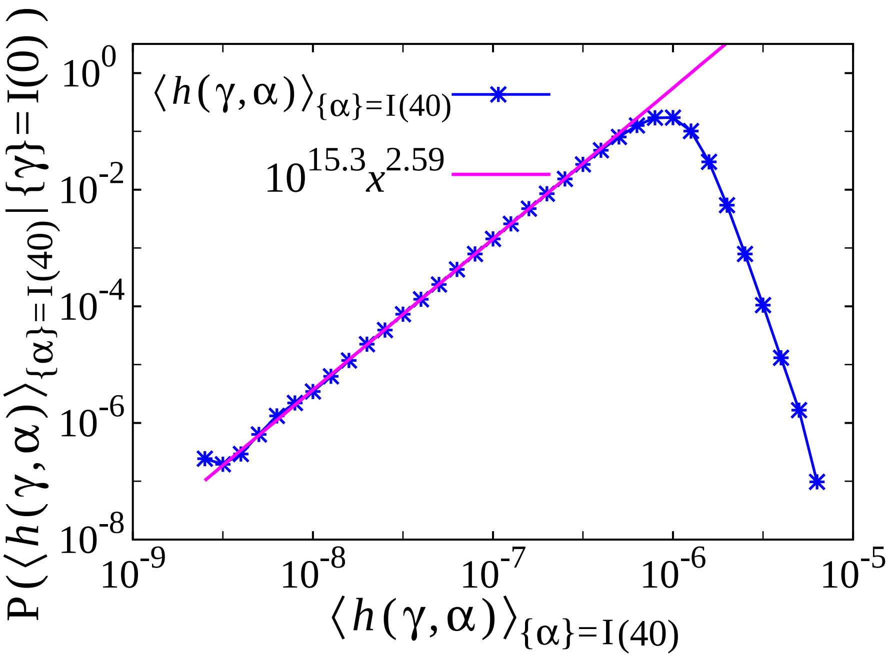


図 3.13: I (0) の個体 γ 毎の 40 世代の個体群 $\{\alpha\}$ = I (40) へ継承する遺産の平均値 $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布。Derrida らの用いた条件で森を 1000 回作成し、それらの結果の平均を取ったものである。直線は得られた分布をその微小領域でフィッティングしたものである。$N = 4096$。

single の結果を見ると、図 3.11 の結果と同様に振動しているように見える。しかし、$N = 32768$ の振幅が $N = 4096$ のものより小さいことからサイズに依るものだと判断できる。また森の数を増やして求めた average ではサイズ依存の効果をある程度取り除くことが出来、S_k は 0.2 に収束することが求められた。図 3.12 より、遺産の分布は 0 でない有限の歪度を持つことが示せた。対数正規分布によく似た図 3.11 や図 3.12 で示される遺産の分布を、本論文では『FS (Finited Skewness) 分布』と呼ぶ。

γ をどう選択しても、血が絶えない限り平均値だけが異なる FS 分布に収束すると考えられる。したがって、0 世代の個体を母集団とした遺産の平均値の分布を求めれば、継承される遺産の分布も決まることになる。γ を母集団とする、I (g) = 40 の個体らに伝える遺産の平均値 $h\,(\gamma, \alpha)$ の分布を両対数プロットで描いた。結果は図 3.13 である。遺産の平均値の分布は作成した森にも依存するため、図 3.13 では森を 1000 回作成して平均を求めた。血が絶えることにより遺産が受け継がれない場合はこの

図には表れず、関係性の場合と同様に 0.2 である。図 3.13 より、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。図 3.13 の直線は $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)} < 10^{-7}$ でフィッテングした関数である。この特徴は先行研究で求められたウェイトに関する結果と類似している。しかし、ウェイトで求めた分布（D 分布）に対して求められたべきの指数はおよそ 1.3 であり、遺産に対して今回求められた指数のおよそ半分であるという違いがある。D 分布と特徴が似ているが指数が異なることから、この分布を『D′ 分布』と呼ぶ。$g > 10$ であれば、微小領域のべき性は観測されるが、遺産の分布が収束する世代数を考慮し、図 3.10 は $g = 40$ とした。

3.2.2 異なる子数分布のモデルの解析

Derrida らの用いた中立モデルにおいて観測された、ウェイトの分布はその値が微小な領域においてべき則を示すという特徴は、子孫方向に木を辿る場合にウェイトに対応する関係性と遺産を用いても観測できた。ただし、ウェイトと関係性のべきの指数と、遺産の指数では 2 倍の差がある。また、ウェイトと関係性は十分世代を隔てた γ—α 間では γ のみに依存する量であるが、遺産は γ だけでなく α にも依存し、その分布は FS 分布に収束する。遺産の量の解析において、子数分布を変えても FS 分布や D′ 分布が得られるかどうかを調べた。検証に用いた子数分布は 3 つで、それぞれ幾何分布、単純な分布 1、2 である。

まず、子数分布が幾何分布となる生物集団を作りだし解析を行った。k 頭の子供を持つ確率 $P_{geometric}(k)$ は

$$P_{geometric}(k) = p(1-p)^k \ (k = 0, 1, 2, \ldots), \tag{3.12}$$

で表される確率分布である。p はパラメータであり、平均子数は $(1-p)/p$ である。水口、堀内らの行った競走馬のデータの解析で、メスの持つ子数の分布は幾何分布に近い形が出ており、この分布も実際の生物の子数と関係しうると考えられる [7, 11]。人口増減が起こらない状況での比較を行うために $p = 1/3$ とし、平均子数が 2 となるようにした。

ポアソン分布での遺産の解析と同様に、γ を変更した時における $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ を求めた。結果は図 3.14 である。図から $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ の値はポアソン分布のモデ

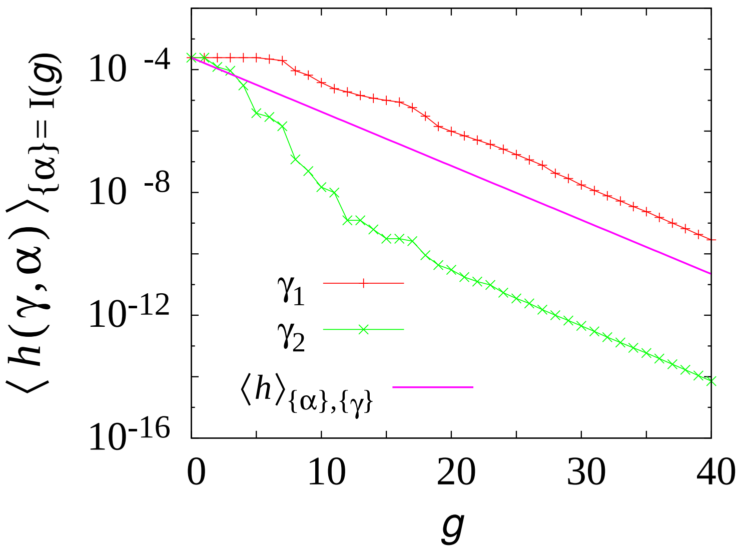


図 3.14: 各世代における遺産の平均値の典型例（γ_1、γ_2）と平均。横軸は世代を表し、縦軸は遺産の平均値を表す。$\langle h\rangle_{\{\alpha\},\{\gamma\}}$ は $g=0$ の個体を母集団とした時の $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の平均である。$N=4096$。

ルと同様 γ に大きく依存する量であることがうかがえる。また十分世代を経ることで $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ は指数関数的に減衰するようになる。その指数は $1-P_{geometric}\,(0)$ であり、子数がポアソン分布の森での指数は $1-P\,(0)$ でこれらは一致する。ただし、十分大きい g 世代までに血が絶えてしまう個体の割合は子数分布がポアソン分布の場合と同様にある有限の値に収束するものの、その大きさが異なる。この割合は Derrida らの用いた非先祖率の収束値と対応する量であるが、子数分布が異なるために幾何分布における非先祖率の漸化式が式 (2.7) に一致しないためである。式 (2.7) は男女の数比が等しく子数がポアソン分布の場合に限るもので、子数分布が幾何分布における非先祖率 $S\,(G)$ は

$$
\begin{aligned}
S_{geometric}\,(G+1) &= \sum_{k=0} P\,(k)\, S_{geometric}\,(G)^k \\
&= \sum_{k=0} p\,(1-p)^k\, S_{geometric}\,(G)^k \\
&= \frac{p}{1-(1-p)\,S_{geometric}\,(G)}\,, \qquad (3.13)\\
S_{geometric}\,(0) &= 1-\frac{1}{N\,(0)}\,, \qquad (3.14)
\end{aligned}
$$

となる。$p = 1/3$ の時、式 (3.13) の $S(G)$ はおよそ 0.5 に収束し、I (0) の中で血が絶えてしまう個体の割合と対応することが確認できた（図 3.15 参照)。

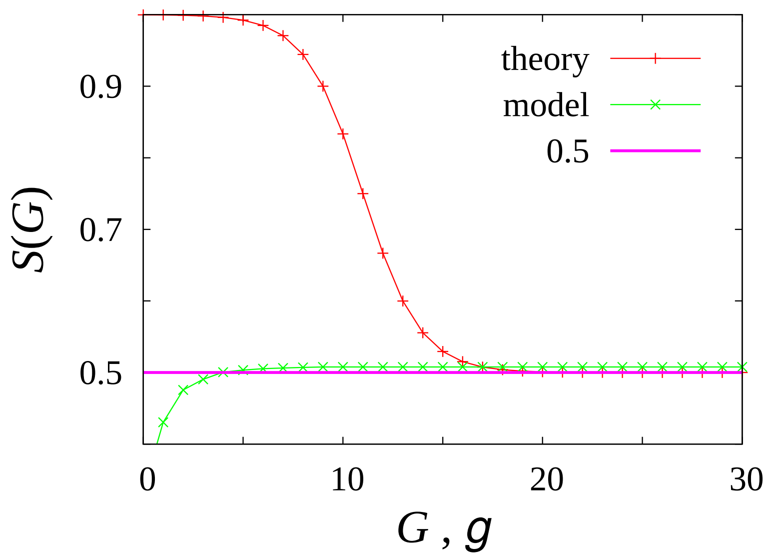


図 3.15: $N = 4096$ における式 (3.13)、式 (3.14) から得られる世代毎の非先祖率と、子数が幾何分布の森から得られた血が絶えてしまう個体の累積分布。横軸は世代を表し、非先祖率においては祖先方向に辿る世代数 G、血が絶える個体の累積分布では子孫方向に辿る世代数 g である。縦軸はそれぞれの割合を表す。theory は式 (3.13) から求めた値を指し、model は指数が幾何分布の森を作成し、I (0) の個体 γ らの中で血が絶えてしまう個体の割合を数値的に求めたもの。

次に、ポアソン分布の時と同様に十分時間の経過した世代における遺産の分布を求めた。その結果が図 3.16 である。γ_1 および γ_2 は図 3.14 の γ_1 と γ_2 にそれぞれ対応している。また、分布の形を比較するために図 3.9、図 3.10 と同様に規格化を行った。幾何分布に変えた場合でも遺産の分布は十分世代を辿ることによって分布の形が収束することが分かった。ただし、得られた分布は図 3.9 で見られた FS 分布と異なり、全体に対し平均値付近のみ高い確率を取る分布の形である。幾何分布のモデルにおける遺産の分布の作成によって、子数分布によって収束する分布の形が異なることが確かめられた。

子数分布が幾何分布のモデルにおいて、0 世代の全個体が 40 世代の個体らに伝える遺産の量を示す分布を両対数プロットで描くと図 3.17 のようになった。遺産の平

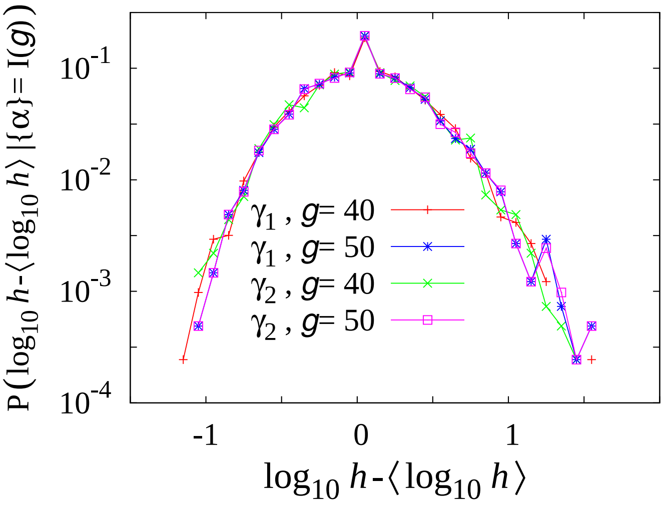


図 3.16: $N = 4096$ の子数が幾何分布の森における遺産の分布。横軸は α らが持つ遺産を表し、縦軸は遺産の平均値を表す。γ_1 および γ_2 は図 3.14 のそれぞれに対応する。

均値の分布は作成したモデルにも依存するため、図 3.13 と同様に森を 1000 回作成して平均を求めた。幾何分布においても $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布の微小な領域においてべき則が見られる。10^{-14}〜10^{-11} の範囲においてフィッティングを試みたところ、べきの指数は 1.75 であった。これは図 3.13 でフィッティングした指数 2.60 と異なり、平均子数が同じであっても子数分布がべきの指数に影響を与えている可能性を示せた。

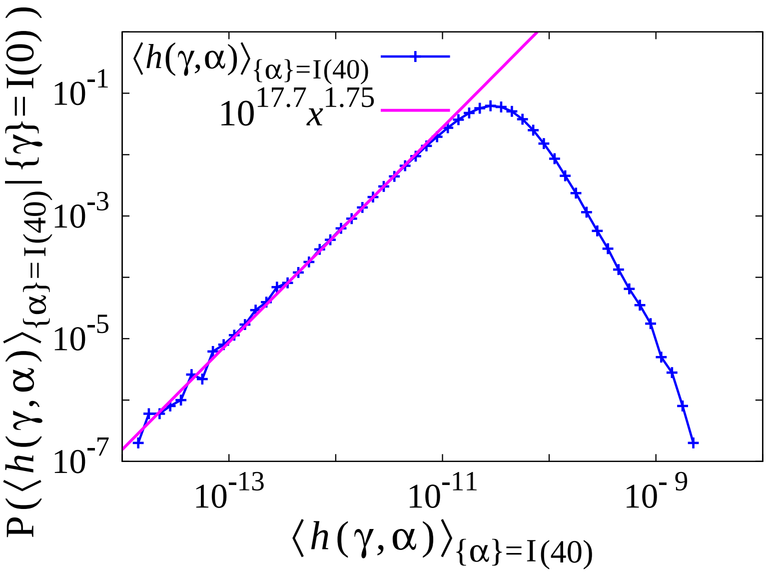


図 3.17: 子数分布が幾何分布の生物集団における、I (0) の個体 γ 毎の 40 世代の個体群 $\{\alpha\} = \mathrm{I}\,(40)$ へ継承する遺産の平均値 $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布。直線は得られた分布をその微小領域でフィッティングしたものである。$N = 4096$。

図 3.17 の結果から、分布の形は子数分布毎に異なると考えられるが、子数分布を変えても $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ の示す分布はその微小領域でべき則を示すと期待される。より子数分布が単純であれば今までに得られた分布を求める定常解や、微小領域において観測されるべきの指数を決定づける要因を見つけやすいと想定される。そこでこの D′ 分布を、子数分布をポアソン分布から非常に単純な分布に変更して解析を行った。用いた子数分布は以下に示す通りで、子数 $P(k)$ を

$$P_{simple\,1}(k) = \begin{cases} p & (k=0) \\ 1-2p & (k=2) \\ p & (k=4) \end{cases} \tag{3.15}$$

と定義した。それぞれが括弧の中の子数が出る確率を示し、p は確率を決めるパラメータである。子数の平均を取ると $0 \times p + 2 \times (1-2p) + 4 \times p = 2$ となり、今まで用いてきたモデルと同様に人口増減が起こらない。この分布を用いて $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布を求めた結果が図 3.18 である。p の値に応じてべきの指数が変化していることが分かる。p は子数分布を決定するパラメータであり、また子数 0 の確率に等しい。べきの指数が子数 0 の確率が等しい別の分布と同じ形を示すかを調べるために、子数分布を式 (3.15) から

$$P_{simple\,2}(k) = \begin{cases} p & (k=0) \\ (1-3p)/2 & (k=1) \\ (1+p)/2 & (k=3) \end{cases} \tag{3.16}$$

へと条件を変え、同様に $\langle h(\gamma, \alpha) \rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布を求めた。その結果べきの指数が 0 の確率のみに依存する数ではないことが確認できた（図 3.19 参照）。

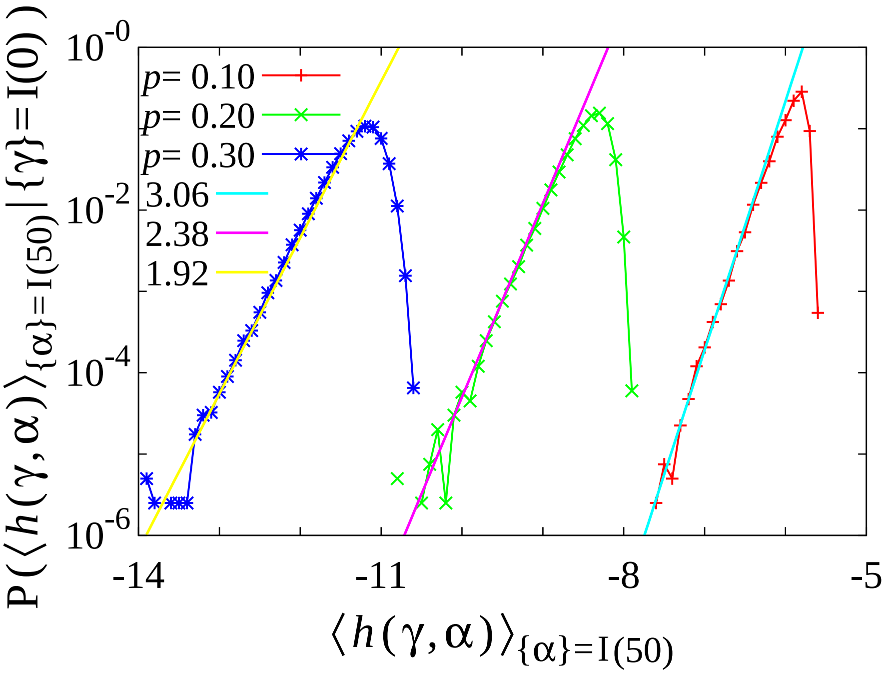


図 3.18: 子数分布が式 (3.15) を満たす森において、$N = 4000$ での I (0) の個体 γ 毎の 50 世代の個体群 $\{\alpha\} = \mathrm{I}\,(50)$ へ継承する遺産の平均値 $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布。生物集団を 1000 回作成し、それらの結果の平均を取った。$p = 0.10,\ 0.20,\ 0.30$ における分布をそれぞれ示しており、直線は得られた分布をそれらの微小領域でフィッティングしたものである。直線の横に書かれているのはフィッティングした直線のべきの指数である。

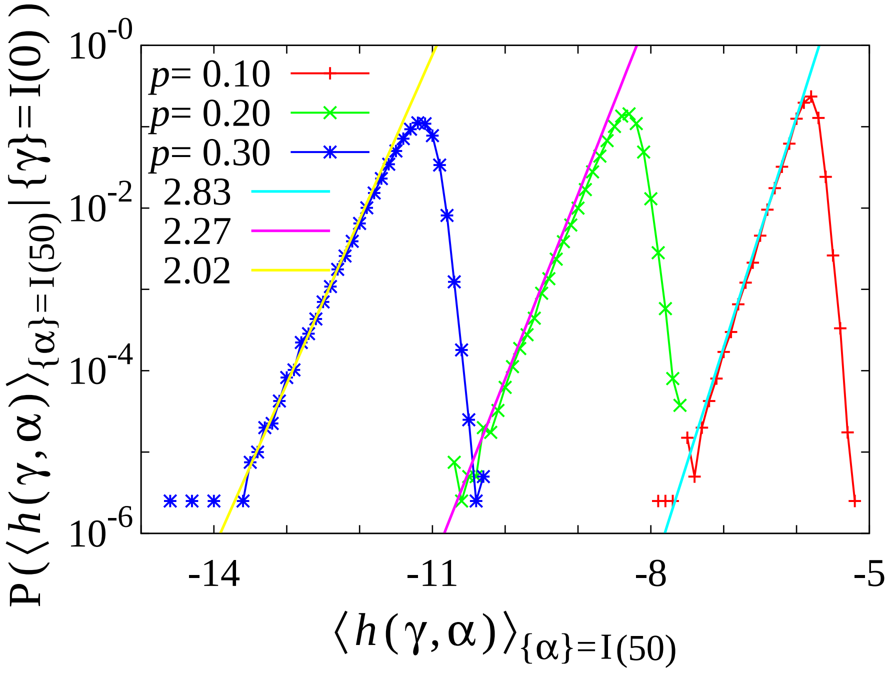


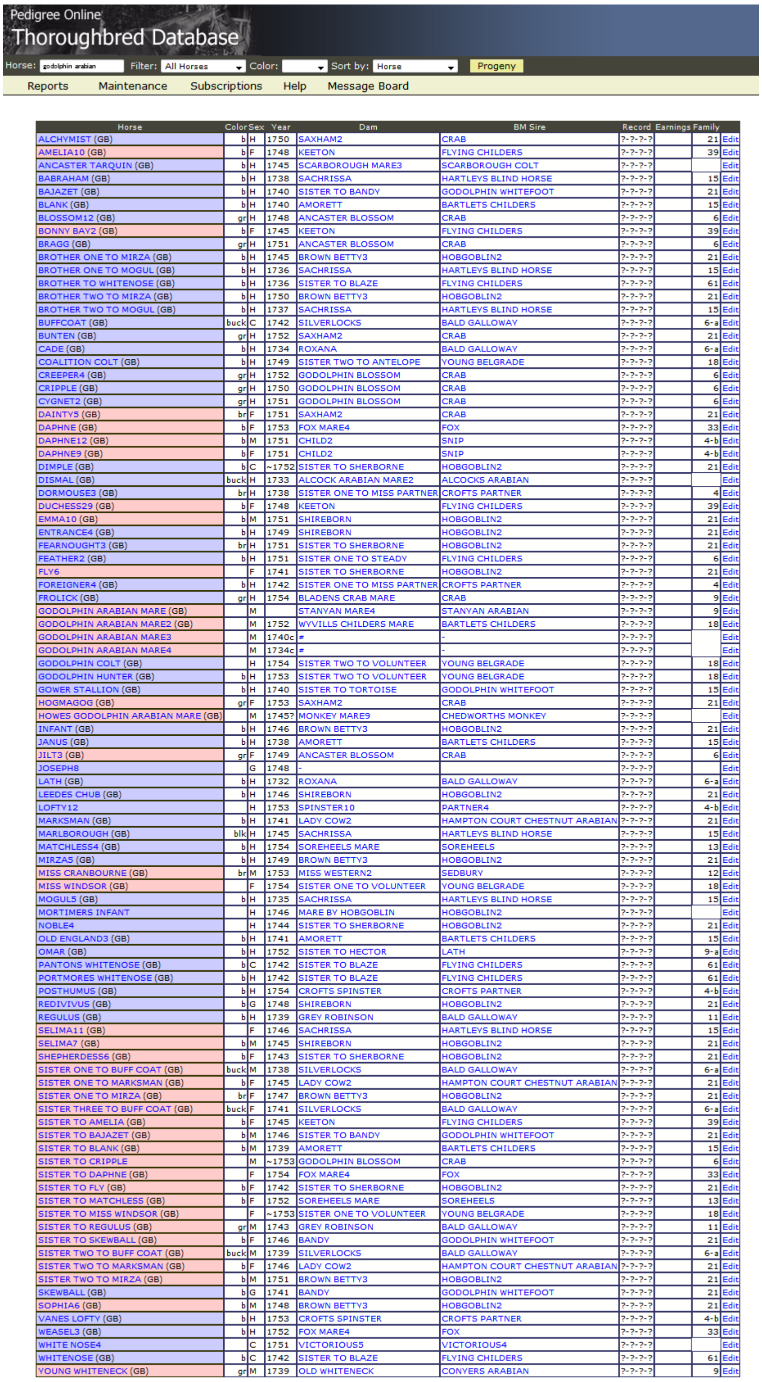
図 3.19: 子数分布が式 (3.16) を満たす森において、$N = 4000$ での I (0) の個体 γ 毎の 50 世代の個体群 $\{\alpha\} = \mathrm{I}\,(50)$ へ継承する遺産の平均値 $\langle\, h\,(\gamma\,,\,\alpha)\,\rangle_{\{\alpha\}=\mathrm{I}(g)}$ の分布。生物集団を 1000 回作成し、それらの結果の平均を取った。$p = 0.10,\ 0.20,\ 0.30$ における分布をそれぞれ示しており、直線は得られた分布をそれらの微小領域でフィッティングしたものである。直線の横に書かれているのはフィッティングした直線のべきの指数である。

第4章　実データの解析

前節で用いた中立モデルは、単純化のために 2.2 節に書かれた 4 つの仮定に基づいている。これらの仮定は、実際の生物集団ではどの程度成立しているのであろうか。また、それらの破れは、家系図の構造にどのような影響を与えるのであろうか。このような疑問に答えるためにも、実際の生物の家系図を調べることには大きな意義があると考えられる。我々は、長期間にわたる木を構築することが出来る実際の生物として競走馬に着目した。解析に用いたデータベース（Pedigree Online Thoroughbred Database）の情報は日々更新されており、現在では 200 万頭を超える競走馬の情報が登録されている [15]。データベース上に存在する競走馬であれば、その名前を検索することで、性別・出生年・子供の名前・配偶馬の名前といった情報を得ることが出来、その結果検索した競走馬の子数も求められる（図 4.1）。検索した競走馬の子供を検索するという作業を繰り返し行うことで、血が途絶えない限り子孫方向への発展を追うことが可能である。競走馬には三大始祖と呼ばれる“ DARLEY ARABIAN ”、“ GODOLPHIN ARABIAN ”、“ BYERLEY TURK ”という競走馬がおり、その三頭の馬から特に“ GODOLPHIN ARABIAN ”に注目した。その馬からデータベース上の馬を子孫方向に辿ることで、家系図上に存在する各馬の名前・性別・出生年・両親の名前・子数を入手し家系図の構造の解析を行った。三大始祖と呼ばれる“ GODOLPHIN ARABIAN ”から辿ることができる子孫は競走馬のデータベースの多くの部分を占めていると考えられ、モデルで用いた一つの γ としての側面を持ちながら、競走馬の森全体の特徴をも示すデータでもあると期待される。

4.1　実データの取得と不完全性

本章では競走馬データから抽出する家系図ネットワークの構造を解析する。このデータは、前章で解析されたモデルと異なり、様々な意味で不完全であり、扱い



Horse	Color	Sex	Year	Dam	BM Sire	Record	Earnings	Family	
ALCHYMIST (GB)	b	H	1750	SAXHAM2	CRAB	?-?-?-?		21	Edit
AMELIA10 (GB)	b	F	1748	KEETON	FLYING CHILDERS	?-?-?-?		39	Edit
ANCASTER TARQUIN (GB)	b	H	1745	SCARBOROUGH MARE3	SCARBOROUGH COLT	?-?-?-?			Edit
BABRAHAM (GB)	b	H	1738	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
BAJAZET (GB)	b	H	1740	SISTER TO BANDY	GODOLPHIN WHITEFOOT	?-?-?-?		21	Edit
BLANK (GB)	b	H	1740	AMORETT	BARTLETS CHILDERS	?-?-?-?		15	Edit
BLOSSOM12 (GB)	gr	H	1748	ANCASTER BLOSSOM	CRAB	?-?-?-?		6	Edit
BONNY BAY2 (GB)	b	F	1745	KEETON	FLYING CHILDERS	?-?-?-?		39	Edit
BRAGG (GB)	gr	H	1751	ANCASTER BLOSSOM	CRAB	?-?-?-?		6	Edit
BROTHER ONE TO MIRZA (GB)	b	H	1745	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
BROTHER ONE TO MOGUL (GB)	b	H	1736	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
BROTHER TO WHITENOSE (GB)	b	H	1736	SISTER TO BLAZE	FLYING CHILDERS	?-?-?-?		61	Edit
BROTHER TWO TO MIRZA (GB)	b	H	1750	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
BROTHER TWO TO MOGUL (GB)	b	H	1737	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
BUFFCOAT (GB)	buck	C	1742	SILVERLOCKS	BALD GALLOWAY	?-?-?-?		6-a	Edit
BUNTEN (GB)	gr	H	1752	SAXHAM2	CRAB	?-?-?-?		21	Edit
CADE (GB)	b	H	1734	ROXANA	BALD GALLOWAY	?-?-?-?		6-a	Edit
COALITION COLT (GB)	b	H	1749	SISTER TWO TO ANTELOPE	YOUNG BELGRADE	?-?-?-?		18	Edit
CREEPER4 (GB)	gr	H	1752	GODOLPHIN BLOSSOM	CRAB	?-?-?-?		6	Edit
CRIPPLE (GB)	gr	H	1750	GODOLPHIN BLOSSOM	CRAB	?-?-?-?		6	Edit
CYGNET2 (GB)	gr	H	1751	GODOLPHIN BLOSSOM	CRAB	?-?-?-?		6	Edit
DAINTY5 (GB)	br	F	1751	SAXHAM2	CRAB	?-?-?-?		21	Edit
DAPHNE (GB)	b	F	1753	FOX MARE4	FOX	?-?-?-?		33	Edit
DAPHNE12 (GB)	b	M	1751	CHILD2	SNIP	?-?-?-?		4-b	Edit
DAPHNE9 (GB)	b	F	1751	CHILD2	SNIP	?-?-?-?		4-b	Edit
DIMPLE (GB)	b	C	~1752	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
DISMAL (GB)	buck	H	1733	ALCOCK ARABIAN MARE2	ALCOCKS ARABIAN	?-?-?-?			Edit
DORMOUSE3 (GB)	br	H	1738	SISTER ONE TO MISS PARTNER	CROFTS PARTNER	?-?-?-?		4	Edit
DUCHESS29 (GB)	b	F	1748	KEETON	FLYING CHILDERS	?-?-?-?		39	Edit
EMMA10 (GB)	b	M	1751	SHIREBORN	HOBGOBLIN2	?-?-?-?		21	Edit
ENTRANCE4 (GB)	b	H	1749	SHIREBORN	HOBGOBLIN2	?-?-?-?		21	Edit
FEARNOUGHT3 (GB)	br	H	1751	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
FEATHER2 (GB)	b	H	1751	SISTER ONE TO STEADY	FLYING CHILDERS	?-?-?-?		6	Edit
FLY6		F	1741	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
FOREIGNER4 (GB)	b	H	1742	SISTER ONE TO MISS PARTNER	CROFTS PARTNER	?-?-?-?		4	Edit
FROLICK (GB)	gr	H	1754	BLADENS CRAB MARE	CRAB	?-?-?-?		9	Edit
GODOLPHIN ARABIAN MARE (GB)		M		STANYAN MARE4	STANYAN ARABIAN	?-?-?-?		9	Edit
GODOLPHIN ARABIAN MARE2 (GB)		M	1752	WYVILLS CHILDERS MARE	BARTLETS CHILDERS	?-?-?-?		18	Edit
GODOLPHIN ARABIAN MARE3		M	1740c	#	-	?-?-?-?			Edit
GODOLPHIN ARABIAN MARE4		M	1734c	#	-	?-?-?-?			Edit
GODOLPHIN COLT (GB)		H	1754	SISTER TWO TO VOLUNTEER	YOUNG BELGRADE	?-?-?-?		18	Edit
GODOLPHIN HUNTER (GB)	b	H	1753	SISTER TWO TO VOLUNTEER	YOUNG BELGRADE	?-?-?-?		18	Edit
GOWER STALLION (GB)	b	H	1740	SISTER TO TORTOISE	GODOLPHIN WHITEFOOT	?-?-?-?		15	Edit
HOGMAGOG (GB)	gr	F	1753	SAXHAM2	CRAB	?-?-?-?		21	Edit
HOWES GODOLPHIN ARABIAN MARE (GB)		M	1745?	MONKEY MARE9	CHEDWORTHS MONKEY	?-?-?-?			Edit
INFANT (GB)	b	H	1746	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
JANUS (GB)	b	H	1738	AMORETT	BARTLETS CHILDERS	?-?-?-?		15	Edit
JILT3 (GB)	gr	F	1749	ANCASTER BLOSSOM	CRAB	?-?-?-?		6	Edit
JOSEPH8		G	1748	-		?-?-?-?			Edit
LATH (GB)	b	H	1732	ROXANA	BALD GALLOWAY	?-?-?-?		6-a	Edit
LEEDES CHUB (GB)	b	H	1746	SHIREBORN	HOBGOBLIN2	?-?-?-?		21	Edit
LOFTY12		H	1753	SPINSTER10	PARTNER4	?-?-?-?		4-b	Edit
MARKSMAN (GB)	b	H	1741	LADY COW2	HAMPTON COURT CHESTNUT ARABIAN	?-?-?-?		21	Edit
MARLBOROUGH (GB)	blk	H	1745	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
MATCHLESS4 (GB)	b	H	1754	SOREHEELS MARE	SOREHEELS	?-?-?-?		13	Edit
MIRZA5 (GB)	b	H	1749	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
MISS CRANBOURNE (GB)	br	M	1753	MISS WESTERN2	SEDBURY	?-?-?-?		12	Edit
MISS WINDSOR (GB)		F	1754	SISTER ONE TO VOLUNTEER	YOUNG BELGRADE	?-?-?-?		18	Edit
MOGUL5 (GB)	b	H	1735	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
MORTIMERS INFANT		H	1746	MARE BY HOBGOBLIN	HOBGOBLIN2	?-?-?-?			Edit
NOBLE4		H	1744	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
OLD ENGLAND3 (GB)	b	H	1741	AMORETT	BARTLETS CHILDERS	?-?-?-?		15	Edit
OMAR (GB)	b	H	1752	SISTER TO HECTOR	LATH	?-?-?-?		9-a	Edit
PANTONS WHITENOSE (GB)	b	C	1742	SISTER TO BLAZE	FLYING CHILDERS	?-?-?-?		61	Edit
PORTMORES WHITENOSE (GB)	b	H	1742	SISTER TO BLAZE	FLYING CHILDERS	?-?-?-?		61	Edit
POSTHUMUS (GB)	b	H	1754	CROFTS SPINSTER	CROFTS PARTNER	?-?-?-?		4-b	Edit
REDIVIVUS (GB)	b	G	1748	SHIREBORN	HOBGOBLIN2	?-?-?-?		21	Edit
REGULUS (GB)	b	H	1739	GREY ROBINSON	BALD GALLOWAY	?-?-?-?		11	Edit
SELIMA11 (GB)		F	1746	SACHRISSA	HARTLEYS BLIND HORSE	?-?-?-?		15	Edit
SELIMA7 (GB)	b	M	1745	SHIREBORN	HOBGOBLIN2	?-?-?-?		21	Edit
SHEPHERDESS6 (GB)	b	F	1743	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
SISTER ONE TO BUFF COAT (GB)	buck	M	1738	SILVERLOCKS	BALD GALLOWAY	?-?-?-?		6-a	Edit
SISTER ONE TO MARKSMAN (GB)	b	F	1745	LADY COW2	HAMPTON COURT CHESTNUT ARABIAN	?-?-?-?		21	Edit
SISTER ONE TO MIRZA (GB)	br	F	1747	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
SISTER THREE TO BUFF COAT (GB)	buck	F	1741	SILVERLOCKS	BALD GALLOWAY	?-?-?-?		6-a	Edit
SISTER TO AMELIA (GB)	b	F	1745	KEETON	FLYING CHILDERS	?-?-?-?		39	Edit
SISTER TO BAJAZET (GB)	b	M	1746	SISTER TO BANDY	GODOLPHIN WHITEFOOT	?-?-?-?		21	Edit
SISTER TO BLANK (GB)	b	M	1739	AMORETT	BARTLETS CHILDERS	?-?-?-?		15	Edit
SISTER TO CRIPPLE		M	~1753	GODOLPHIN BLOSSOM	CRAB	?-?-?-?		6	Edit
SISTER TO DAPHNE (GB)		F	1754	FOX MARE4	FOX	?-?-?-?		33	Edit
SISTER TO FLY (GB)	b	F	1742	SISTER TO SHERBORNE	HOBGOBLIN2	?-?-?-?		21	Edit
SISTER TO MATCHLESS (GB)	b	F	1752	SOREHEELS MARE	SOREHEELS	?-?-?-?		13	Edit
SISTER TO MISS WINDSOR (GB)		F	~1753	SISTER ONE TO VOLUNTEER	YOUNG BELGRADE	?-?-?-?		18	Edit
SISTER TO REGULUS (GB)	gr	M	1743	GREY ROBINSON	BALD GALLOWAY	?-?-?-?		11	Edit
SISTER TO SKEWBALL (GB)	b	F	1746	BANDY	GODOLPHIN WHITEFOOT	?-?-?-?		21	Edit
SISTER TWO TO BUFF COAT (GB)	buck	M	1739	SILVERLOCKS	BALD GALLOWAY	?-?-?-?		6-a	Edit
SISTER TWO TO MARKSMAN (GB)	b	F	1746	LADY COW2	HAMPTON COURT CHESTNUT ARABIAN	?-?-?-?		21	Edit
SISTER TWO TO MIRZA (GB)	b	M	1751	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
SKEWBALL (GB)	b	G	1741	BANDY	GODOLPHIN WHITEFOOT	?-?-?-?		21	Edit
SOPHIA6 (GB)	b	M	1748	BROWN BETTY3	HOBGOBLIN2	?-?-?-?		21	Edit
VANES LOFTY (GB)	b	H	1753	CROFTS SPINSTER	CROFTS PARTNER	?-?-?-?		4-b	Edit
WEASEL3 (GB)	b	H	1752	FOX MARE4	FOX	?-?-?-?		33	Edit
WHITE NOSE4		C	1751	VICTORIOUS5	VICTORIOUS4	?-?-?-?			Edit
WHITENOSE (GB)	b	C	1742	SISTER TO BLAZE	FLYING CHILDERS	?-?-?-?		61	Edit
YOUNG WHITENECK (GB)	gr	M	1739	OLD WHITENECK	CONYERS ARABIAN	?-?-?-?		9	Edit

図 4.1: データベースから検索した”GODOLPHIN ARABIAN”の子供のリスト。

には注意が必要である。まずは、典型的なデータを見てみよう。競走馬の名前はアルファベットで登録されているが、同名もあり得る。同名がいる場合は、図 4.1 の“ GODOLPHIN ARABIAN MARE ”にあるように、名前に数字を付加することで個体を識別するための名前の唯一性を確保している。

次に不完全な点を列挙してみよう。まず、データベースが閉じていない。先祖を辿っていくと遅かれ早かれ名前の分からない親に辿りつき、その以前の家系図は辿れない。最も過去にまで遡れたのは出生年が 1612 年の“ FAMILY NUMBER SIX ”であるが、それ以外の祖先は全てそれ以降で家系図を遡れなくなる。モデルでも最古の世代より遡ることが出来ないが、競走馬のデータの場合、名前の分からない親が出てくる世代は様々である。たとえば 2010 年に生まれた“ LITTLE ANGEL8 ”でも、その父親は“ unknown ”である。これについては 4.3 節で述べる。

次に、データベースは日々更新している。新しく生まれたものと思われる個体のみならず、古い時代のデータも書き換えられることがある。本研究で使用されたデータは 2013 年 11 月 6 日〜2014 年 1 月 15 日の間に取得されたデータである。ただし、その時点で登録されている全てのデータを取得できていないことにも注意しなければならない。

また、データには欠損や矛盾がある。出生年が不明な個体や、親もしくは子の出生年と矛盾があるようなデータもある。例えば、図 4.1 の“ GODOLPHIN ARABIAN MARE ”は出生年が不明である。このような馬は現在探索が終わっている個体約 140 万頭の中で 1 万頭登録されている。また、親より子の出生年が早い個体や、親との年齢差が 100 歳以上離れている個体も存在する。

他にも生まれた子供が全てデータベースに登録されているわけではないと思われる。というのは生まれた子が全て競走馬になるわけではないからである。

このように、競走馬のデータは不完全であり 100 %の信頼度がおけるわけではない。しかしながら、保存されたそのデータベースは膨大であり、我々はその親子関係を追跡することで個体数 1403702 頭、240 年を超える長大な家系図を構築することに成功した（図 4.2 参照）。これは質・量ともに類を見ないものであり、不完全なことを差し引いても解析すべき貴重なデータである。図 4.2 は三次元プロットであり、青い丸がオスを表し、赤い三角がメスを表す。薄い青色の部分は親子関係を結んだ線を表しているが、数が膨大で入り乱れているため線ではなく面に見えている。

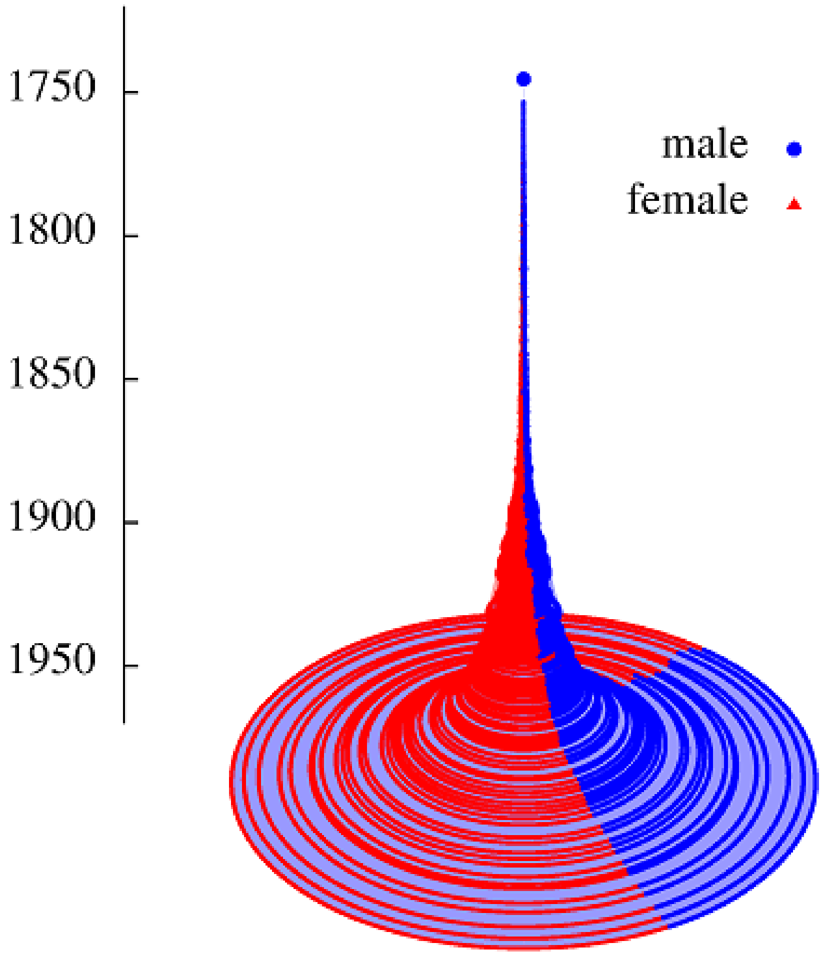


図 4.2: “ GODOLPHIN ARABIAN ”から構築した 240 年分の家系図。薄い青が親子関係を表す線。縦軸が競走馬の出生年を表す。

4.2 世代の定義づけ

ウェイトや関係性・遺産といった量は、局所的な親子関係に基づいて定義されているので、競走馬の家系図でもそのまま定義可能である。一方、その平均値や分布の世代依存性を考える時には、世代を適切に定義しなければならないが、これは自明ではない。というのも、今まで解析してきた中立モデルでは起こらなかった世代の重複を考慮する必要があるからである。例えば図 4.3 を見てみよう。図 4.3 の中の丸と三角の印はそれぞれオスとメスを表す。また、縦軸が出生年を表しており、上にある印の個体ほど過去に生まれたことを表す。黄色で示した主個体 γ に対して複数の役割を有する子孫個体 α が、パス 1（図中左の矢印）では二世代しか離れていないが、パス 2（図中右の矢印）では四世代離れている。これは競走馬の生殖年齢に有限の幅があることが原因である。生殖年齢の幅についての詳細は後述するが、競走馬において親子間の年齢差は 3〜30 歳ほどある。このように、複数のパスで異なる距離をとる場合、個体間の相対的な世代は唯一には定まらない。ところが、前章で述べたウェイトや関係性・遺産の世代依存性は複数のパスが存在するというネットワーク構造を反映した結果であり、これらと比較するためには、各個体に対して世代をユニークに決める必要がある。

ここでは、各個体に主個体に対する世代を一意的に決める方法を二種類紹介する。一つは最短世代である。これは注目する γ の世代を 0 とし、γ から α に到達するまでに通過する馬が一番少なくなる場合に辿る馬の数を定義とする。図 4.3 を用いて最短世代の取り方を具体的に説明しよう。例えば黄色で示した個体を γ とし、緑色で示した子孫個体 α の世代を定めると、γ の子供は直接辿ることができるので子供の世代数は 1 となり、γ の孫は γ とその子供を通過して辿りつくため子供の世代数より 1 つ大きい 2 となる。図 4.3 の矢印で示したように辿るパスが複数考えられる個体の場合は経由する馬の数が最少となるように選ぶため、γ からメスの子供を辿るパスを選択しその個体の最短世代は 2 となる。基本的に子の世代は親の世代より 1 つ大きくなる点が中立モデルの世代の取り方と等しいと言える。また競走馬のデータベースには親子関係がはっきりしているが出生年が定かでない個体が存在しており、そうした個体に対しても一意に世代を定められるのがこの設定方法の利点である。しかし出生年を無視した取り方であるために、図 4.3 の例で示したように子の世代が親の世代を上回ることなどがあり、世代数と γ との続柄が必ずしも一致しな

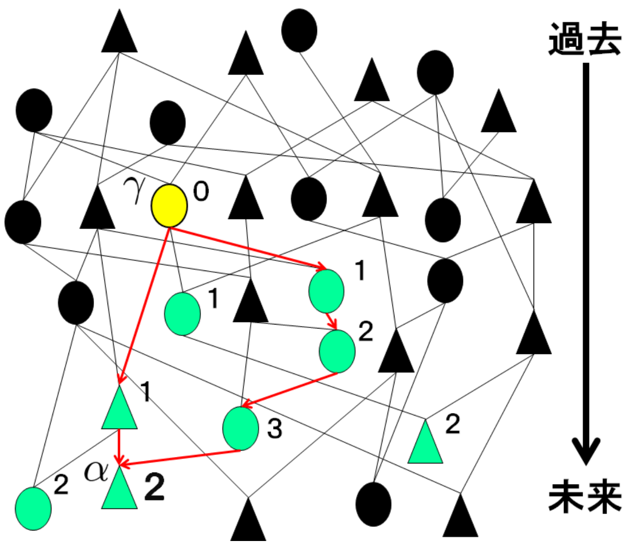


図 4.3: 世代の重複を加味した家系図ネットワークの模式図と最短世代。丸で示した個体をオス、三角で示した個体をメスとする。縦方向が出生年を表す。黄色で示した個体が注目する γ を表し、緑色で示した個体は γ の子孫個体 α らを表す。γ と α の右上に書かれている数字は最短世代を表す。

いことに注意が必要である。また、子孫方向にのみ辿っても辿りつかない個体には世代が定義されず、大域的な定義ではない。

二つ目は年代世代である。一定の期間内に生まれた個体は全て同じ世代であるとする考え方である。年代世代の取り方の例を図 4.4 に示す。この図は世代の取り方以外は先ほどの図 4.3 と同じ記号を用いている。年代世代は、例えば赤線で示したように家系図上の個体を特定の期間で区切り、主個体 γ が生まれた期間を世代 0、次を 1 ⋯ と世代を定める。出生年に応じて昇順に並ぶ点がモデルの世代と等しい。年代ごとに世代を設定しているため最短世代のように親子間で世代数が逆転するといったことは起こらず、一般的に世代という言葉が有する性質に反しない。しかし競走馬のデータには出生年不明の個体が存在しているため、それらの個体を年代世代では記述できないという短所がある。また、図 4.4 の子孫個体 α_1 と α_2 にあるように、区切る期間によっては親子関係にあるにも関わらず同じ世代に属することや、γ の子 α_3 が世代 2 にいるといった親子の世代差が 1 より大きいことも起こり、こちらも世代数が γ と子孫らの続柄を完全には表現できないことに注意が必要である。しかし、出生年に基づいているので、”GODOLPHIN ARABIAN”の子孫でなくても定

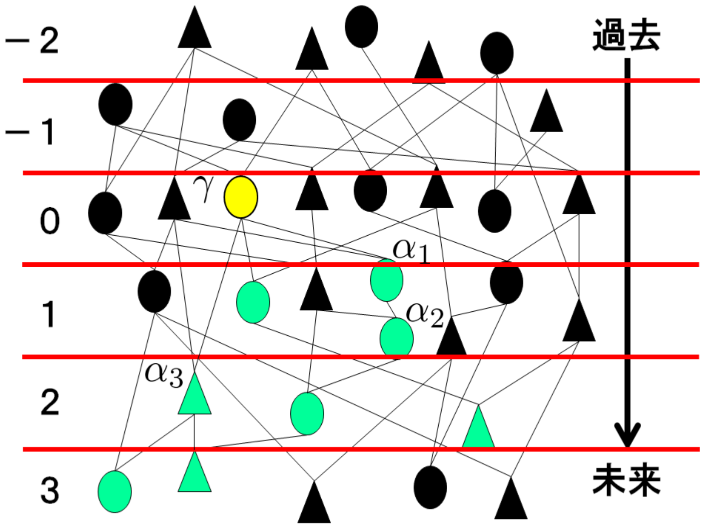


図 4.4: 図 4.3 の家系図ネットワークに対する年代世代。図中の左の数が年代世代における世代数を表す。

義できる。その意味では、大域的な定義と言えよう。競走馬の解析においては、年代世代の期間を十年間とし、“ GODOLPHIN ARABIAN ”の出生年が 1724 年であるため、1720〜1729 年生まれの個体を年代世代 0 として解析を行った。

他にも、世代の取り方としては主個体 γ から子孫 α への全てのパスに対してその長さの平均を定義とした平均関係世代や、もっとも長いパスの長さを定義とした最長関係世代なども考えられるが、探査方法からくる制限によりここでは採用しない。詳細は次節で述べる。

4.3 有効なデータの範囲

以降の解析に用いられたデータは、”GODOLPHIN ARABIAN”の子孫であるが、データベースに登録されたものの一部に限定されている。というのは、競走馬のデータベース（Pedigree Online）には膨大な数の競走馬が登録されており、その全てを検索し尽くしていないからである。検索方法については本章の始めに簡単に述べたが、各競走馬を調べた際に得られるその子供の名前を辿り続ける、言い換えると最短世代の昇順に競走馬を検索している。そのため、図 4.3 の例にもあるように最短世代が同じ個体であっても出生年に開きがあり、ある子孫の出生年が 2013 年だとし

ても、その子孫の情報を得た時点で取り出せている情報で 2013 年までに生まれた“ GODOLPHIN ARABIAN ”の子孫個体を全て取り終えているとは限らない。現在情報を取り終えているデータは最短世代で 16 世代までであり、17 世代を検索中である。取得中のデータがどの程度の年まで取り終えているかを計るために、各最短世代毎に年代世代を用いて頭数分布を求めた。その結果が図 4.5 と図 4.6 である。これらの図は最短世代で g_M だけ検索した時点の競走馬の頭数分布である。例えば、$g_M = 1$ は最短世代が 1 の個体の頭数分布、$g_M = 2$ は最短世代が 1 と 2 の個体の頭数分布となる。$g_M \geq 4$ から 1900 年以降に分布が飛んでいるが、これはデータベースに幾つかの間違いがあるためと考えられる。実際、データベースの中には親よりも出生年が早い馬がのべ 163 頭確認できており、親と年齢差が 50 を超える馬もデータベースにはのべ 443 頭存在する。しかし、全体の個体数に対してそうしたエラーは少なく、以降の結果に大きな影響を与えないと考えられるため、誤りだと考えられる個体も母集団に含めたまま解析を進めた。図 4.5 を見ると、最短世代の異なる個体が同じ年代世代を取ることから、世代が重複していることが確認できる。現時点では 16 世代までは検索が終了しているため、17 世代の途中のデータと 16 世代までのデータの差分を取ることでおおよそどの年度まで取り終えているかを計ることが出来る。図 4.6 を見ると 1960 年（1960〜1969 年）までは $g_M = 17$ と $g_M = 16$ の検索頭数のグラフは重なっている。17 世代が検索途中であるため多少の誤差は考えられるが、現在、大部分に取得出来ている馬のデータは 1969 年までであると結論付けた。

モデルと同様の現象が起こると仮定すると、ある個体の子孫はその集団全体に広がっていく。したがって、子孫方向に木を構築していくと、ある世代より後に産まれた子孫個体の個体数はその世代の個体数と一致すると期待される。すなわち、その世代からのデータは木の構造を捉えながら森全体の構造を捉えることに等しくなる。ある年代世代に属する個体数と、“ GODOLPHIN ARABIAN ”の木の中の個体であり、かつその年代世代に属する個体数が一致しているかどうかを測る指標として、両親個体が“ GODOLPHIN ARABIAN ”の木の中に属するかどうかを調べた。もし“ GODOLPHIN ARABIAN ”の木に属する個体数が、ある世代から森の個体数と一致すれば、その後の年代の馬は“ GODOLPHIN ARABIAN ”の木の中にいる個体が親になるためである。“ GODOLPHIN ARABIAN ”の木の中の個体に対し、両親が

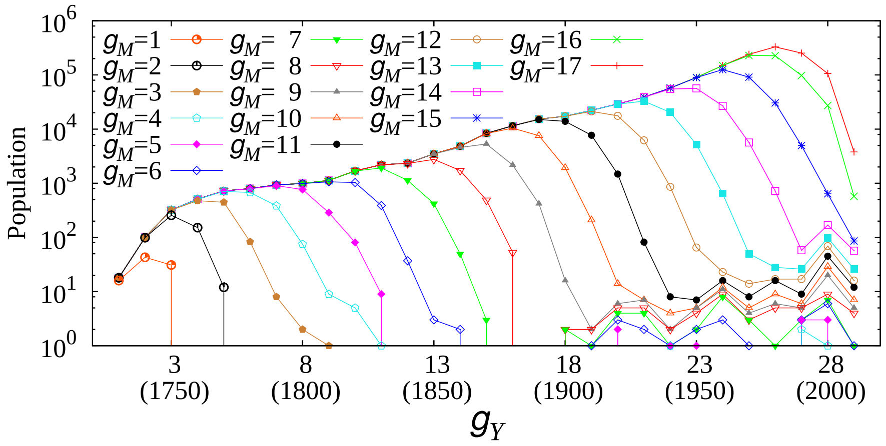


図 4.5: “ GODOLPHIN ARABIAN ”を先祖とする木の子孫個体 1392451 頭のデータにおける、最短世代毎の検索した競走馬の累積分布。横軸が年代世代を表し、縦軸は頭数の片対数を表す。$g_M = 1$〜17 がそれぞれその世代まで累積の頭数から求めた分布。頭数は 10 年毎にまとめてあり、その期間の始めの年に書かれている（例えば 1900〜1909 年の馬のデータは 1900 年として記載）。

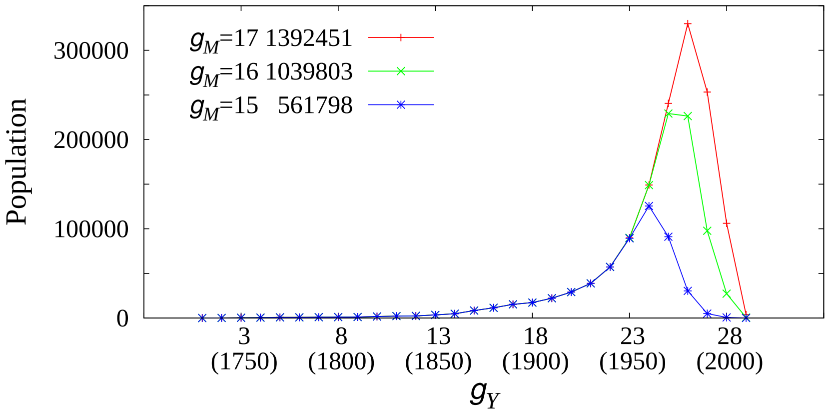


図 4.6: “ GODOLPHIN ARABIAN ”を先祖とする木の子孫個体 1392451 頭のデータにおける、最短世代毎の検索した競走馬の累積分布。横軸が年代世代を表し、縦軸は頭数を表す。$g_M = 15$〜17 がそれぞれその世代まで累積の頭数から求めた分布。世代の隣の数字はその世代までに検索できた競走馬の頭数。図 4.5 同様頭数は 10 年毎にまとめてあり、その期間の始めの年に書かれている。

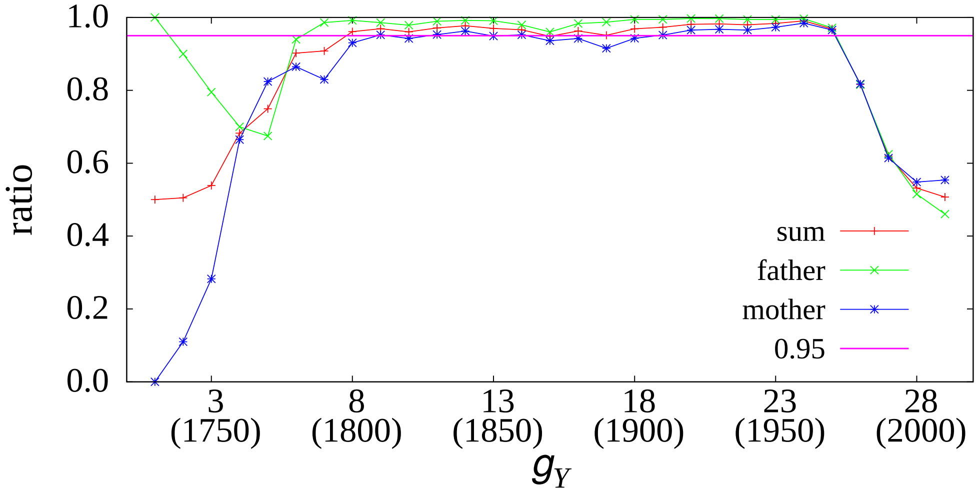


図 4.7: “ GODOLPHIN ARABIAN ”を先祖とする家系図上の子孫個体 1392451 頭のデータにおける、年代世代毎の両親が“ GODOLPHIN ARABIAN ”の木に属する割合。横軸が年代年代を表し、縦軸は木に属する競走馬の割合を表す。

“ GODOLPHIN ARABIAN ”の木の中の個体にいる割合を求めた。その結果を図 4.7 に示す。横軸は年代世代を表し、縦軸は両親が“ GODOLPHIN ARABIAN ”の木に属している割合である。“ father ”はその年代に属する“ GODOLPHIN ARABIAN ”の子孫個体の中で父馬が判明している個体の割合、“ mother ”は母馬が判明している個体の割合で、“ sum ”は父馬が判明している個体数と母馬が判明している個体数の和を、その年代の“ GODOLPHIN ARABIAN ”の子孫の個体数の 2 倍で割ったものである。始めは“ GODOLPHIN ARABIAN ”(1724 年生まれのオス馬) の子供に限られるため、“ father ”が 1.0、“ mother ”が 0.0 である。その後、子孫個体の増加に伴って、両者ともほぼ一定の割合に漸近する。性別を無視すれば 1800 年以降の個体は 95 %以上の両親が“ GODOLPHIN ARABIAN ”の子孫である。1980 年以降の割合が減少しているが、これは図 4.5、4.6 で示した検索終了年代の境界より後の世代である。したがって、競走馬のデータが検索中であり、親子関係がまだ確認できていない個体が存在することが原因であると考えられる。親子関係が確認できていないことで、1980 年以降の個体は関係性や遺産の継承が正しく計算されていないと考えられる。このため、年代世代においては 1800 年から 1969 年までの期間が、関係性や遺産の分布、加えて競走馬全体の家系図の特徴を見出す上で有効なデータだ

と仮定し以降の解析を進めた。

同様に、最短世代に対しても関係性や遺産の分布を測定することを考える。分布を求める際に適切な世代の範囲は、最短世代において両親が“ GODOLPHIN ARABIAN ”の木に属する割合によって決める。その結果を図 4.8 に示す。横軸は最短世代を表し、縦軸は両親が“ GODOLPHIN ARABIAN ”の木に属している割合である。主個体がオスであるため、$g_M = 1$ の“ father ”の値は 1.0 を示す。$g_M = 1$ の“ mother ”の値が 0 でないが、これは $g_M = 1$ に属する個体の幾つかが、図 4.3 で示した例のように、“ GODOLPHIN ARABIAN ”とその子孫個体で g_M が自身以上のメス馬を母親に持つためである。$g_M > 1$ では、年代世代と同様に、子孫個体の増加に伴って、両者ともほぼ一定の割合に漸近する。性別を無視すれば $g_M = 6$ 以降の世代の個体は 95 %以上の両親が“ GODOLPHIN ARABIAN ”の子孫である。$g_M = 16, 17$ において割合が減少しているが、これはまだ調べ終えていない $g_M = 17$ 以降の世代に片親を持つためだと考えられる。このため、$g_M = 16$ 以降の世代では関係性や遺産の継承が十分に行われておらず、これらの分布を求めることは適当ではないと考えられる。このため、最短世代では $g_M = 6 \sim 15$ において関係性や遺産の分布の特徴を見出す上で有効なデータだと仮定し以降の解析を進めた。

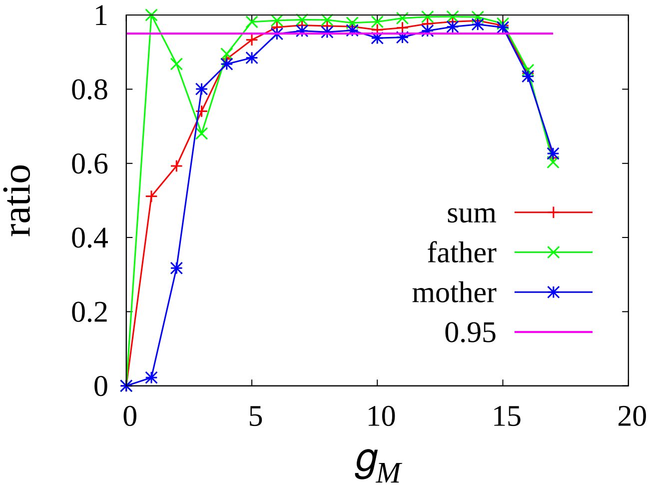


図 4.8: “ GODOLPHIN ARABIAN ”を先祖とする家系図上の子孫個体 1403702 頭のデータにおける、最短年代毎の両親が“ GODOLPHIN ARABIAN ”の木に属する割合。横軸が最短年代を表し、縦軸は木に属する競走馬の割合を表す。

年代世代、および最短世代において、世代を経ることでいずれもほぼ全ての親が“ GODOLPHIN ARABIAN ”の木に属していることが確認できたが、100 %にはならない。この原因は二つ考えられる。一つは“ GODOLPHIN ARABIAN ”の木が集団全体に広がりきっていないためである。もう一つは 4.1 節で述べた未登録の馬すなわち外部からの流入である。両者の寄与を見積るために、我々は、年代世代において世代毎の“ unknown ”な親の割合を求めた。結果は図 4.9 である。横軸は年代世

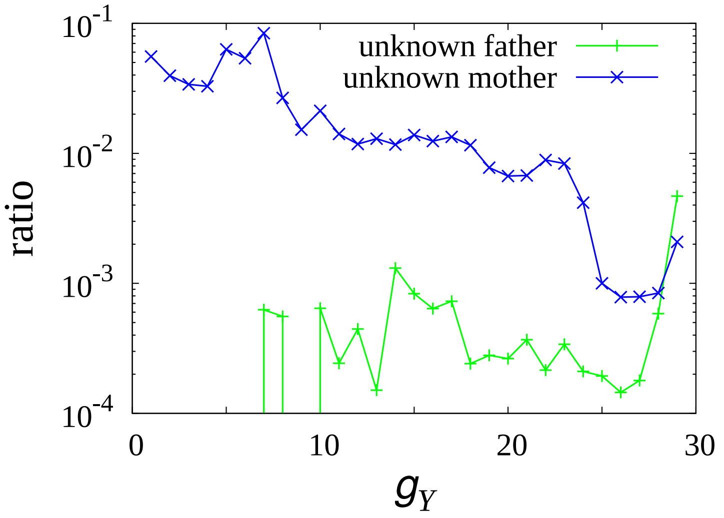


図 4.9: “ GODOLPHIN ARABIAN ”を先祖とする家系図上の子孫個体 1392451 頭のデータにおける、年代世代毎の両親がデータベース上に存在しない割合。横軸が年代年代を表し、縦軸は“ unknown ”な親の割合を表す。

代、縦軸は各年代世代での個体数に対する“ unknown ”な親の割合を表す。“ father ”はその年代に属する“ GODOLPHIN ARABIAN ”の子孫個体の中で、父馬がデータベース上に存在しない割合を表し、“ mother ”はその年代に属する“ GODOLPHIN ARABIAN ”の子孫個体の中で、母馬がデータベース上に存在しない割合を表す。これを見ると、0.1 %以下の割合ではあるが、データベース外の競走馬が父親となっており、また 1 %程度の割合でデータベース外の競走馬が母親となっていることが分かる。このように、競走馬のデータベースにおいてはモデルでの解析と異なり生物集団が閉じていないことにも注意しなければならない。

4.4　子数分布

γ を始祖とする子孫方向に辿る木の中で、出生年が判明している 1800〜1969 年の競走馬約 45 万頭から、オスとメスの子数分布を求めた。本節では、それぞれの子数分布や、それらの平均や分散をまとめた表を記す。

まず、オスの子数分布について述べる。45 万頭の競走馬のうち、オスは 140521 頭であり、その子数分布を図 4.10 に示す。図 4.10 の横軸は子数を表しており、縦軸

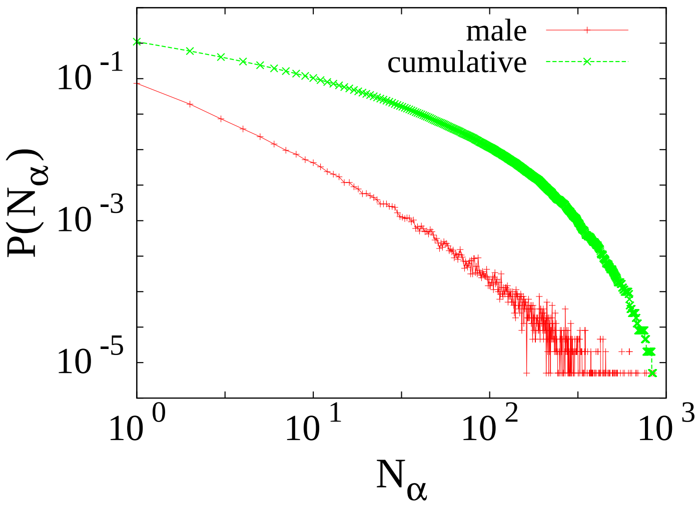


図 4.10: “ GODOLPHIN ARABIAN ”を先祖とする家系図上の子孫個体 140521 頭のデータから求めたオスの子数分布。横軸が子数を表し、縦軸は確率分布を表す。male は確率分布、cumulative は N_α 以上の子供を持つ割合を示す累積頻度分布を表す。$N_\alpha = 0$ の個体の割合は 0.67 である。

は確率を表している両対数プロットである。赤で示された“ male ”は通常の度数分布、緑で示された“ cumulative ”は累積頻度分布を表している。子数が 0 の割合は 0.67 程度である。累積頻度分布は、べき的な特徴を度数分布よりよく測定できる分布関数で、そのべきの指数は度数分布のべきの指数より 1 大きい数となる。子数分布で単一のべきを示していれば、累積頻度分布でも同様に単一のべきで表すことが出来る。水口、堀内の行った先行研究 [7, 11] ではオスの子数分布をカットオフを考慮したべき的な分布と仮定しており、我々が求めた子数分布も一見するとべきを示しているように見える。ただし、図 4.10 で累積頻度分布は曲線を描いているため、

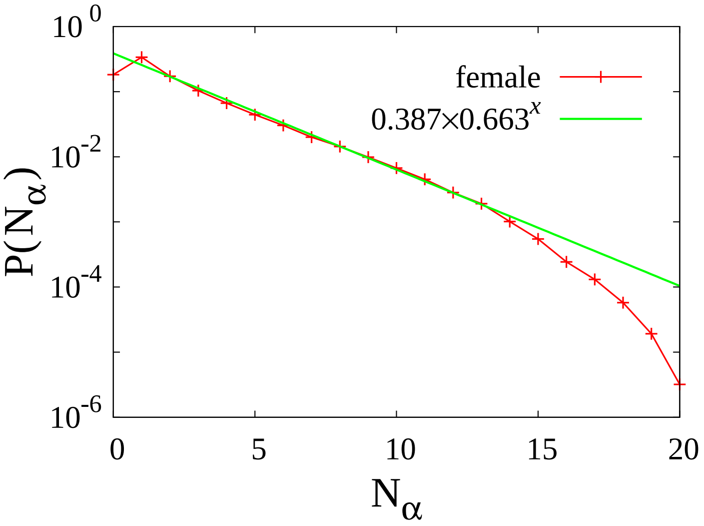


図 4.11: “ GODOLPHIN ARABIAN ”を先祖とする家系図上のメス 312825 頭のデータから求めた子数分布の片対数プロット。横軸が子数を表し、縦軸は確率分布を表す。female は確率分布を表し、緑の実線は N_α が 1 から 13 の範囲でフィッティングした関数を表す。

先行研究で仮定された単一のべきを用いた子数分布は、競走馬のオスの子数分布を表現する点で不十分であったと言える。

次にメスの子数分布を図 4.11 に示す。メスは 312825 頭である。図 4.11 は片対数プロットであり、図中の緑実線は子数 1〜13 の範囲を直線でフィッティングしたものである。オスは 6 割が子を持たない一方で 100 頭以上子を持つ個体もおり、その子数分布は分散が大きい。これに対してメスの子数は 20 まででありオスに比べて分散が小さい。これは、メスが受精から出産までの約一年間は生殖活動が行えないこと、また競争馬の寿命が 30 歳ほどであることに起因すると考えられる。図 4.11 から、子数が 2〜12 の範囲では指数関数的に減衰していると考えられる。しかし、P (0) や P (1) がフィッティングした関数からずれていること、また子数が 13 から落ち込みが始まることから、競走馬の森をよりよく再現するためには、幾何分布におけるフィッティングは改良の余地があると考えられる。

雌雄が異なる子数分布に従っていることが示され、中立モデルの仮定の一つが破れていることが確認できた。ここでは、それぞれの子数分布の平均や、雌雄の数比について表 4.1 にまとめた。表 4.1 から、雌雄の数比が等しいという仮定についても

	合計	オス	メス
子数の平均	3.19	5.36	2.21
子数の分散	181	567	5.42
標準偏差	13.5	23.8	2.33
数比	1.00	0.31	0.69

表 4.1: 出生年が 1800〜1969 年までの個体から求めた平均値、分散、標準偏差、数比。

	合計	オス	メス
子数の平均	3.0	7.1	2.7
数比	1.0	0.4	0.6

表 4.2: 先行研究 [11] により見積られた子数分布の平均値と数比。

破れていることが確認できる。オスの平均子数は時折現れる子数の巨大な個体の影響を受けるためか、1800〜1969 年までの競走馬の g_Y 毎の平均子数にはばらつきがみられるが、同じ期間のメスの g_Y 毎の平均子数の差は 2.1 ± 0.2 程度に抑えられている。そのため長期間をまとめてみると平均子数や分布には違いがないと判断し、森全体の特徴として上述のように期間を取った。また、先行研究と本研究で求めた雌雄の子数分布は、大まかには同じ特徴がみられるが、両者の平均子数と雌雄の数比は異なる（表 4.2 参照）。競走馬の森では世代の重複が起きているため、オスの平均子数とオスの数比の積は、メスの平均子数とメスの数比の積と一致しなくとも誤りではない。しかし、先行研究におけるこれらの積はそれぞれ 2.8 と 1.6 と大きく異なるのに対し、本研究におけるこれらの積は 1.66 と 1.52 でありこの偏りの差が小さい。先行研究における、子数分布を求める際に用いた競走馬のサンプル数はオス 1256 頭、メス 3093 頭であり [6, 11]、数比を求める際に用いたサンプル数はオス 17925 頭、メス 30435 頭であった [11]。本研究で用いたサンプル数は 45 万頭でありその数に大きな違いがあることから、先行研究での見積もりで得られた二つの積が大きく異なるのは、サンプル数の少なさやその偏りが原因の一つとして考えられる。

平均子数 m を用いて世代毎の子孫増加率を表すことを考える。世代の重複の無い生物集団の時、役割の重複が無ければ次世代の子孫数は最大で m 倍になると考えら

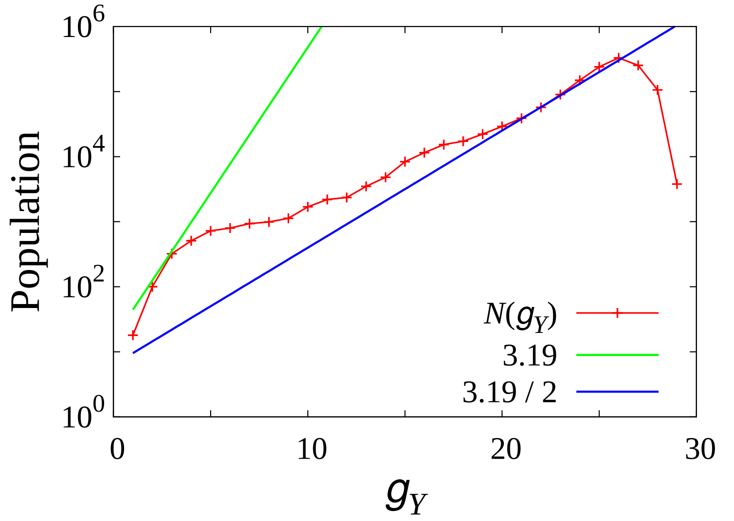


図 4.12: 図 4.5 の $g_Y = 17$ における累積頭数分布と、平均子数から予想される子孫数の最大増加率と最小増加率の直線。横軸が年代世代を表し、縦軸は頭数の片対数を表す。緑と青の実線は、それぞれ出生年が 1800〜1969 年の競走馬における平均子数と親との平均年齢差から求めた、予想される最大増加率と最小増加率を示す直線である。

れる。逆に、最小の子孫増加率となるのは次世代の子孫の両親が今の世代の子孫に限られる場合で $m/2$ 倍である。“ GODOLPHIN ARABIAN ”の子孫増加率はおおよそこの二つ数の間を取ると考えられる。そこで、図 4.5 の $g_Y = 17$ における頭数分布に、本節で求めた平均子数を用いた最大の子孫増加率を描く直線と、平均子数の半分を用いた最小の子孫増加率を描く直線を重ねた（図 4.12 参照）。横軸と縦軸は図 4.5 と等しい。緑と青の実線は平均子数から予想される子孫数の最大増加率と最小増加率を示したものである。具体的にこれらの実線の指数を求めるには、一世代が何年分であるかを求めなければならない。年代世代の期間を便宜上 10 年と定めたため、実際の競走馬における親子間の年齢差は 10 年ではない。本論文では、この一世代と定める年齢差を、平均子数を求める際に用いた競走馬の親子間の平均年齢差で求めた。求められた 11.25 年を一世代分として定義し、今まで解析で用いてきた 10 年分の期間に換算した。したがって、年代世代において予想される子孫数の最大増加率は $3.19^{10/11.25} \approx 2.80$ であり、最小増加率は $(3.19/2)^{10/11.25} \approx 1.05$ である。それぞれ図 4.12 の緑と青の実線の指数に相当する。緑の実線と頭数分布とを比べる

と、年代世代における子孫増加率が平均子数 3.19 を上回る期間は $g_Y \leq 2$ に限られることが分かる。そのため、“ GODOLPHIN ARABIAN ”の木では早くから役割の重複が始まっていると考えられる。

図 4.12 の青の実線と頭数分布とを比べると、年代世代における子孫増加率が、平均子数の半分 3.19/2 を上回る期間はおよそ $g_Y = 22$ 以降に限られることが分かる。また、$g_Y = 5$ あたりから $g_Y = 22$ までの長い間、子孫増加率が 3.19/2 を下回ることも見て取れる。一世代が適切に取れていれば、直観的には子孫増加率が 3.19/2 を下回ることは少ないと考えられる。したがって、子孫増加率が 3.19/2 を下回る一因は真の一世代が適切に取れていないと考えられ、それは次節で述べる。

4.5 親子間の年齢差

この節では年代世代、および最短世代を用いた親子間の平均年齢差の推移と、年齢差の分布を紹介する。競走馬の親子間の年齢差は、ある一つの定数ではなく個体毎に異なる。

まずは年代世代で見てみよう。年代世代を用いた、父母との平均年齢差は図 4.13 である。“ father ”と“ mother ”はそれぞれ父馬・母馬との平均年齢差を表し、“ sum ”は性別を無視して求めた親子の平均年齢差である。横軸の値は年代世代を表し、下の括弧は一世代と定めた期間の始めの年を表す。ただし、”GODOLPHIN ARABIAN” の木に属している個体の間のみで年齢差を求めているため、木に属さない親との間の年齢差は勘定されていない。始めの数世代が低い値を取っているが、これは“ GODOLPHIN ARABIAN ”の出生年が $g_Y = 0$ (1724) 年であり、“ GODOLPHIN ARABIAN ”やその子を親に持つ $g_Y = 3$（1750 年）頃までの馬は高い値を取らないためである。これは親子間の年齢差は 3〜30 ほどであり、主個体やその子供の出生年から 30 年以上経った世代からでないと年齢差の上限が抑えられ平均値は小さくなる傾向がある。例えば、$g_Y = 1$ の個体は 1724 年生まれの主個体か $g_Y = 1$ に生まれた 1730〜1739 年生まれの別の個体としか親子間の年齢差を計算していない。そのため年齢差の上限が 15 年に抑えられる。

逆に $g_Y = 25$ から急激に平均年齢差が増加しているが、これは現在のデータが検索途中であるためと考えられる。$g_Y = 25$ 以降の馬のデータは探索しきっておらず、

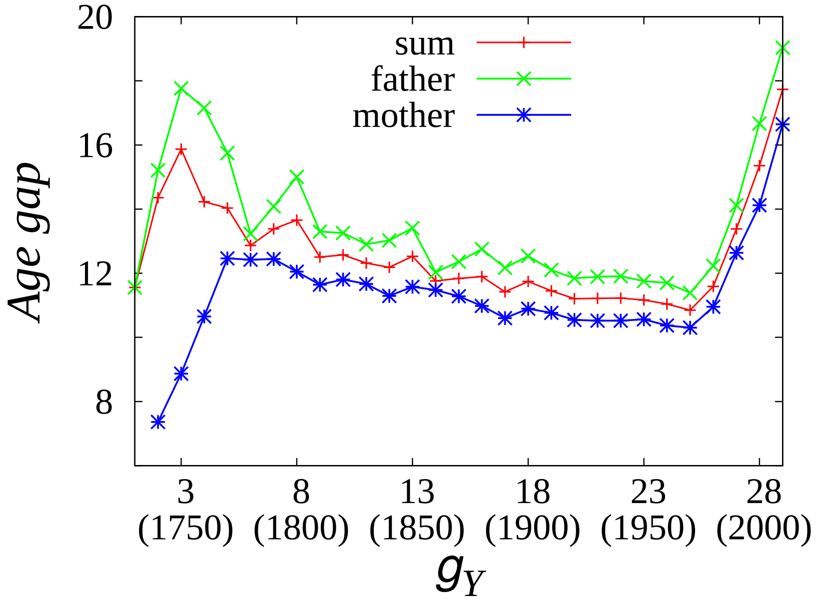


図 4.13: 年代世代における、親子間の平均年齢差の推移。横軸が年代世代、縦軸が年齢差を表す。

かつ馬の探索方法が最短世代の昇順であることから、現在調べ終えている $g_Y = 25$ 以降の馬は年齢差が大きい個体に偏っていると考えられる。

図 4.13 において、父母と子の年齢差は減少している傾向が見られる。この原因は繁殖技術の向上により、生殖可能な年齢が低下した可能性が考えられる。また、競走馬の本来の寿命は 20〜25 歳程度と言われているが、種馬として育てられる一部の馬を除き、その年齢に達することなく殺処分されてしまい高齢で子をなすことが難しい環境である。このことも親子間の年齢差が減少している一因と考えられる。また、年齢差の平均が父馬と母馬で 1 歳程度離れていることは、先行研究で示されている結果と合致する [7, 11]。平均年齢差の緩やかな減少が起きていることにより、ある定数を用いて実際の一世代と見る方法は適当ではないと言える。前節の図 4.12 では $g_Y = 22$ までの子孫増加率が 3.19/2 を下回っていたが、図 4.13 の結果と照らし合わせることで、その原因は g_Y 毎の親子間の平均年齢差が 11.25 年を上回っていたことが一因であると分かった。

長期間を母集団にして平均年齢差を求めても、年代世代の人口増加を上手く表現できないことが分かった。では、図 4.13 で求まった各年代世代毎の平均年齢差と、

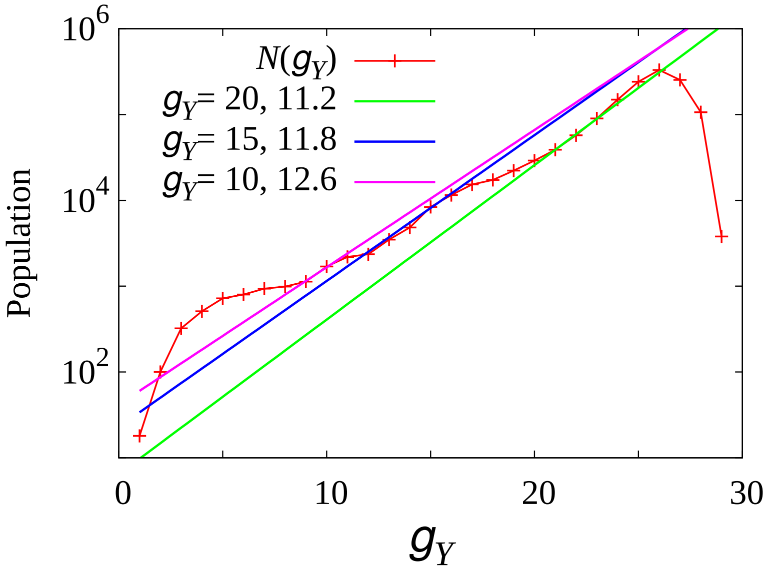


図 4.14: 年代世代毎の頭数分布。$N\,(g_Y)$ は図 4.5 の $g_Y = 17$ に等しい。実線の隣に書かれた数字は、それぞれ図 4.13 から得られた $g_Y = 10,\ 15,\ 20$ の平均年齢差を表す。

予想される最小子孫増加率 3.19/2 を用いれば年代世代の人口増加を表現できるだろうか。その結果を図 4.14 に示す。横軸が年代世代、縦軸が個体数を表す。直線の指数は、各 g_Y での平均年齢差と最少子孫増加率 3.19/2 を用いて、年代世代の定義である 10 年間での子孫増加率に換算した値である。これらの直線と各 g_Y 付近の傾きを見比べると、いずれも非常に近い値を取ることが分かった。

世代の取り方による違いを比較するため、最短世代においても同様に父母との年齢差を求めた。結果は図 4.15 である。最短世代が $g_M = 1$ の個体は親がオスの“ GODOLPHIN ARABIAN ”のみであるため、“ mother ”のデータは $g_M = 2$ からとなっている。$g_M = 1$ の“ father ”の年齢差が非常に高い数値を表しているが、これは“ GODOLPHIN ARABIAN ”の子数が 95 頭と多く、生涯にわたって子を残し続けたためと考えられる。“ mother ”のデータは $g_M = 4$ まで上昇しているが、これは木が広がりきる前のデータの偏りであると考えられる。$g_M = 4$ 以降では父母ともに年齢差は減少傾向を示している。また、データを取得中の 17 世代においては年代世代で見られたような急激な年齢差の上昇は見られない。年代世代で年齢差を求めると、父母との年齢差に加え自身の出生年にも依ってしまう。そのため、その制約のない最短世代では平均を求める際に用いられる個体数が増えることで、年齢差の

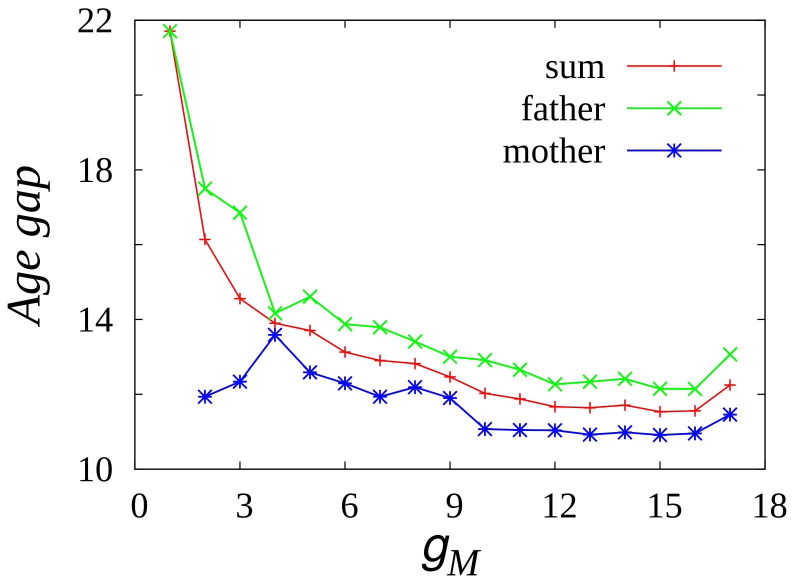


図 4.15: 最短世代における、親子間の平均年齢差の推移。横軸が最短世代、縦軸が年齢差を表す。

偏りが抑えられたのではないかと考えられる。また、最短世代においても、父馬との年齢差と母馬との年齢差には 1 歳程度の差がみられる。

年代世代における母馬との平均年齢差の最低値は $g_Y = 27$ での 10.3 歳、最短世代では $g_M = 15$ での 10.9 歳であるように、世代の取り方の違いにより平均年齢差の値は異なる。しかし、どちらの取り方においても世代を経ることで親子間の年齢差が減少すること、父馬との平均年齢差と母馬との平均年齢差には 1 歳ほどとなる特徴が見られた。

平均値については図 4.13、4.15 となったが、父母との年齢差の分布はどのようになっているであろうか。その結果を図 4.16、4.17、4.18、4.19 に記す。これらは順に、年代世代・最短世代における父馬との年齢差分布、年代世代・最短世代における母馬との年齢差分布を表す。横軸は子数、縦軸は確率を表す。 これらの分布を見ると、 4.2 節で前述したように競走馬においては親子間の年齢差にばらつきが見られ、かつ父母によって平均年齢差が異なることからも世代の重複が起きることが分かる。また、この図 4.16〜4.19 から、年代・最短世代および性別に関わらず、親子の年齢差が大きい確率は減少し、親子の年齢差が小さい確率はわずかではあるが増

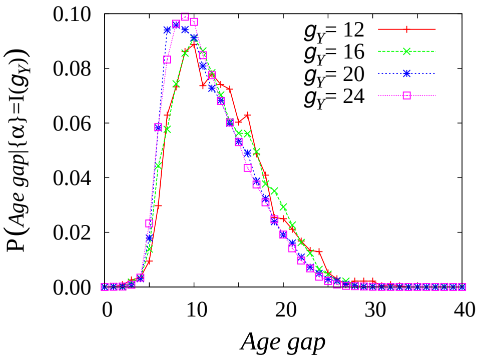


図 4.16: 年代世代における、父馬との年齢差分布。横軸が年齢差、縦軸が確率を表す。

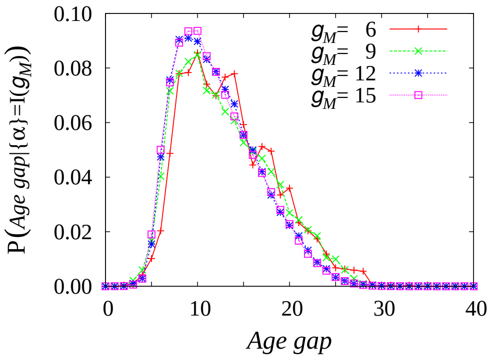


図 4.17: 最短世代における、父馬との年齢差分布。

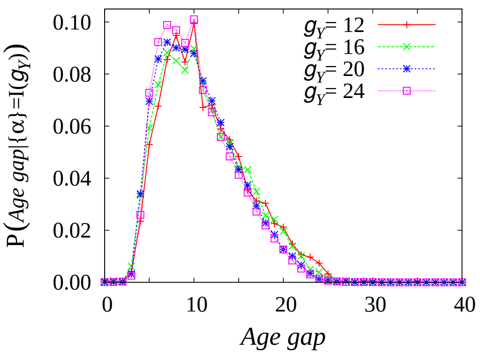


図 4.18: 年代世代における、母馬との年齢差分布。

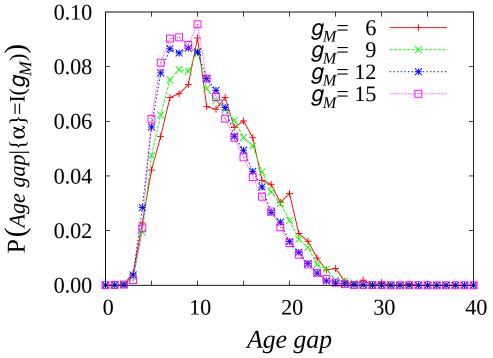


図 4.19: 最短世代における、母馬との年齢差分布。

加していることが分かる。平均年齢差の緩やかな減少傾向はこの分布の変化を表していたと言える。父母による分布の形に大きな差は見られず、平均値が異なる以外は同じ特徴を持つ分布だと考えられる。

4.6 関係性

“ GODOLPHIN ARABIAN ”の木について、モデルとの比較を行うため関係性の平均値・分布を求めた。ただし、本節及び次節で述べる分布の母集団はデータベースに登録されている全ての馬を対象としたものではなく、“ GODOLPHIN ARABIAN ”から子孫方向にのみパスを辿った場合に現れる個体に限定される。その理由は最短世代では“ GODOLPHIN ARABIAN ”の木に属しない個体に対して世代を定義できないからである。年代世代においてはこの限りではないが、$g_Y = 8$ からはデータベース上の大部分の個体を含んでいると考えられ、これ自体でもモデルの結果と比較できると期待される。そのため、最短世代・年代世代ともに“ GODOLPHIN ARABIAN ”の木に属する個体を母集団として解析を行った。

まず、関係性の平均値を解析した。ただし、母集団が異なるためこの値で見られる特徴はモデルの解析で見られた特徴とは異なる。一般的に、世代の重複のないモデルかつ役割の重複が起こらない場合において、関係性の定義から役割の重複が起きない限り α の持つ関係性は 2^{-g} で記述される。そのため平均値は役割の重複が起きない限り、2^{-1} の指数で指数関数的に減衰する単調減少な量である。役割の重複が起こることでいずれ子孫個体は 2^{-g} より大きい関係性を持ち平均値の減衰は抑えられ、子孫個体の大部分の両親がともに関係性を持つ一主個体の子孫が集団全体に広がる一ことである一定の値に収束する。集団全体に広がることで母集団がその世代の個体全てになるため、その時点からは母集団が 3.1 節での解析と一致する。世代の重複がある森とでは条件が異なるが、個々の持つ関係性は両親個体のみに依存するため大まかな特徴は変わらないと考えられる。以上を踏まえ、競走馬における関係性の平均値の年代世代依存性を求めたところ図 4.20 のようになった。横軸は年代世代を表し、縦軸は関係性の平均値を表す。年代世代において $g_Y = 5$ までは関係性の平均値は一定の指数で指数関数的に減衰しており、その値は 0.75 である。0.5 より大きな値を示しており、これは役割や世代の重複が起きたことによる効果であると考えられる。$g_Y = 8$ からはある一定の値に収束しているように見えることから、この図においても“ GODOLPHIN ARABIAN ”の木に両親個体の大部分がその木に属していることが読み取ることができる。$g_Y = 25$ から値が減少しているが、図 4.7 で得られた結果から、これはデータを取得し終わっていないことにより、実際には存在する親子関係の中に探査されていないものが残されているためである。最短世代

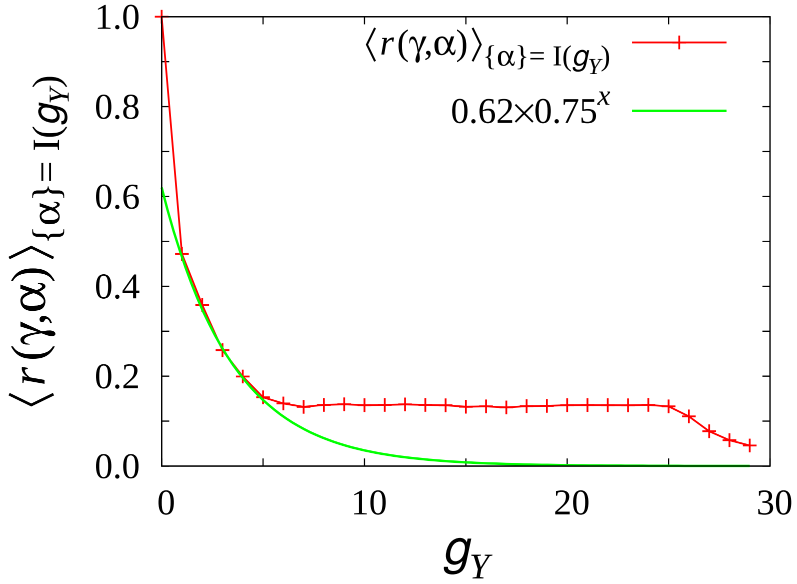


図 4.20: 年代世代における、関係性の平均値の推移。横軸が年代世代、関係性の平均値を表す。

においても同様の解析を行ったが、図 4.5 での図の結果から予想されるように、世代の取り方が異なるために世代数やそれぞれの数値は異なるが、世代を経ることによって見られる特徴が変わらなかった。

$g_Y = 8$、$g_M = 6$ 以降の世代については、競走馬における母集団がデータベース上の集団全体に行き渡っているであろうと考えられるため、モデルの結果と直接比較可能である。$g_Y = 8$、$g_M = 6$ 以降の世代について関係性の分布を求めた。年代世代における結果は図 4.21、4.22 である。横軸は関係性の値、縦軸は確率の対数プロットを表す。分布の変化を見るため、平均値からのずれをプロットした。また、メッシュを 0.01, 0.04 でとりそれぞれ関係性の度数分布を求めた。これらを見ると、モデルの結果とは異なり分布は非対称性が強く、正規分布で近似するのは適切ではないことがわかる。最短世代においても同じような特徴が得られたため、ここでは年代世代での結果のみに触れる。Derrida らの中立モデルの仮定である、世代が重複しない・雌雄の数比が等しい・ランダムに配偶するという条件を満たせば、関係性の分布は正規分布で近似できると言える。しかし、競走馬の森において世代は重複しており、また雌雄の数比は異なる。加えて競走馬はより良い成績を残した馬やその

子孫がより多くの子を残す傾向にあると考えられる。それらの効果により、分布が正規分布になっていないと考えられる。図 4.21 を見ると、この分布は世代が経つにつれて平均値付近と 0.05 にピークを持つ分布に収束しているのではないかとも考えられる。しかし、メッシュを 0.04 と粗くとり分布を求めると、2 つのピークを持つ分布の特徴は失われるが、分布に現れる指数関数的な傾きが減少し続けているように見える。そのため、関係性の分布がまだ過渡過程である可能性を否定できず、今後新たな馬のデータを得ることでこの分布は現在予想される形とは別の形に収束するとも考えられる。

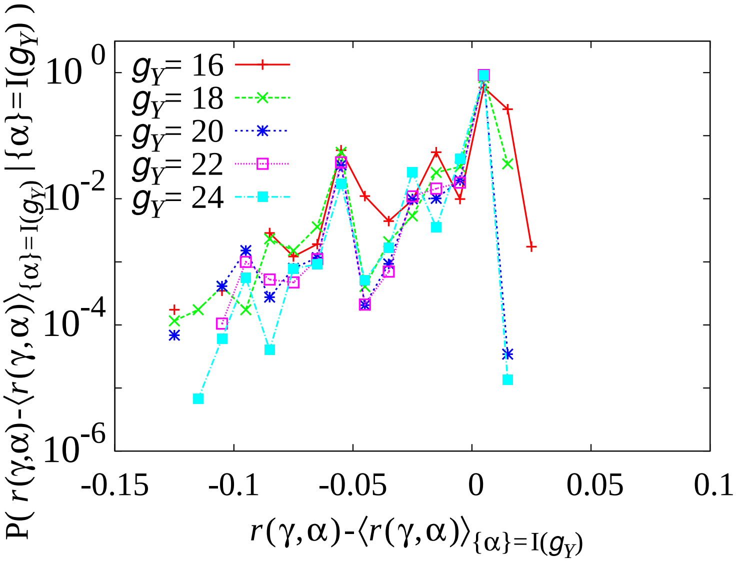


図 4.21: 年代世代における関係性の分布の世代依存性。メッシュは 0.01。

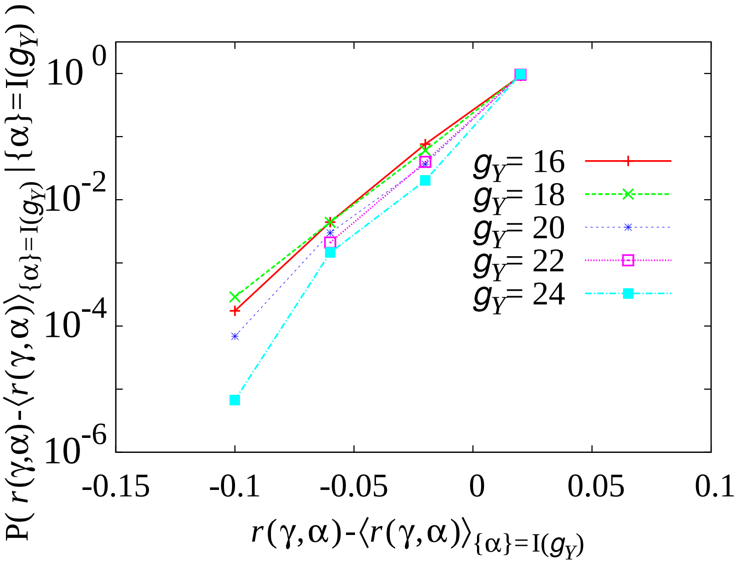


図 4.22: 年代世代における関係性の分布の世代依存性。メッシュは 0.01。

4.7 遺産

この節では、遺産の平均値と遺産の分布について述べる。前節の関係性同様、“ GODOLPHIN ARABIAN ”の子孫でできた木を母集団にした場合、子孫が競走馬の集団全体に行き渡る前と後ではその性質が少し異なる。遺産の平均値は子数 0 の子孫個体が現れることで減少し、また各世代での遺産の総量が初期値を上回ることがないため、母集団に属する個体数が増加することでも減少する量である。集団全体を母集団とした場合、木に属する個体を母集団とした場合のどちらもこの特徴は等しいが、母集団を木に属する個体に限定すると g_Y が小さい範囲では個体数が集団全体に対して非常に少なくなるため、集団全体を母集団にするよりも遺産の平均値は高い値を示す。世代を経ることで木に属する個体数と集団の個体数が一致してからは母集団が等しくなる。したがって、3.2 節での結果と比較できるのは、木が集団全体に広がってからということになる。遺産の平均値の年代世代依存性を図 4.23 に示す。横軸が年代世代、縦軸が遺産の平均値を表す。緑の実線は $g_Y = 15$ から $g_Y = 25$ までの区間を直線でフィッティングしたものである。$g_Y = 27$ から遺産の値が上昇しているが、これは図 4.13 から平均年齢差が大きい個体、すなわちほかの個体よりも遺産の大きい親から遺産を受け継いでいる個体が多いためだと考えられる。遺産の平均値は $g_Y = 27$ までは単調減少しており、$g_Y = 13$ から $g_Y = 25$ はある一定の指数で指数関数的に減衰しているように見える。この期間はほぼデータベース全体の個体数であると考えられる $g_Y = 8$ 以降であるため、モデルにおける図 3.8、図 3.14 の $\langle h \rangle_{\{\alpha\},\{\gamma\}}$ に対応するものである。この指数は 0.53 であり、単位年当たりに換算すると 0.94 となる。平均年齢差は減少傾向にあるが、これを無視し出生年が 1800〜1969 年の個体を母集団ととり、父母問わず平均年齢差を求めると 11.25 年となる。これを真の競走馬における一世代だと仮定すると、一世代毎に減衰する遺産の平均値の割合は 0.49 となる。中立モデルにおける、平均子数 m の生物集団において遺産の平均値の減衰する指数は $2 \cdot (1 - \mathrm{P}(0)) / m$ で現せるため、表 4.1 の平均子数・雌雄の数比、雌雄の $\mathrm{P}(0)$ の値を用いて指数を求めたところ 0.42 という値が得られた。世代の重複を無視した計算ではあるが、平均年齢差・子数分布・雌雄の数比といった情報である程度まで遺産の平均値の減衰を表現できることが分かった。

次に最短世代における遺産の平均値の世代依存性を図 4.24 に示す。横軸は最短世代、縦軸は遺産の平均値を表す。$g_M = 8$ から $g_M = 16$ までの区間を直線でフィッ

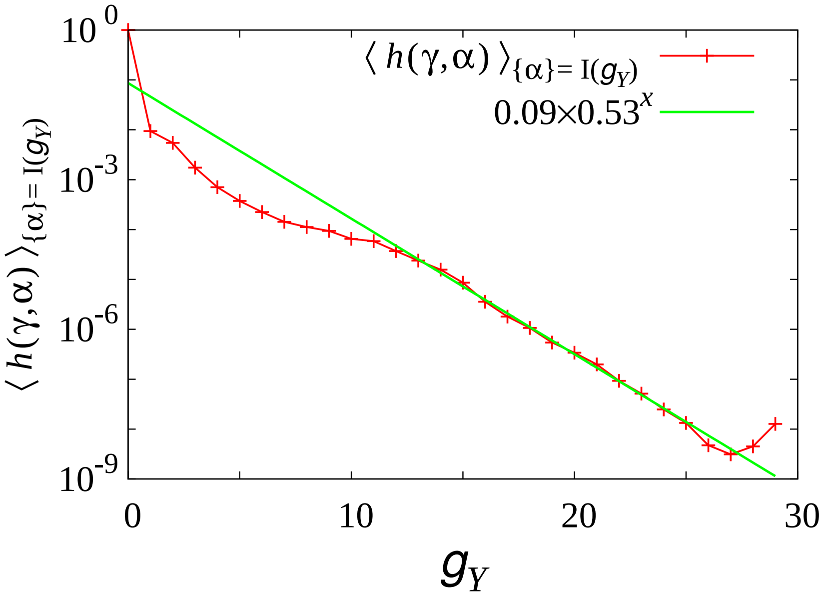


図 4.23: 年代世代における、遺産の平均値の世代依存性。横軸が年代世代、縦軸が遺産の平均値を表す。

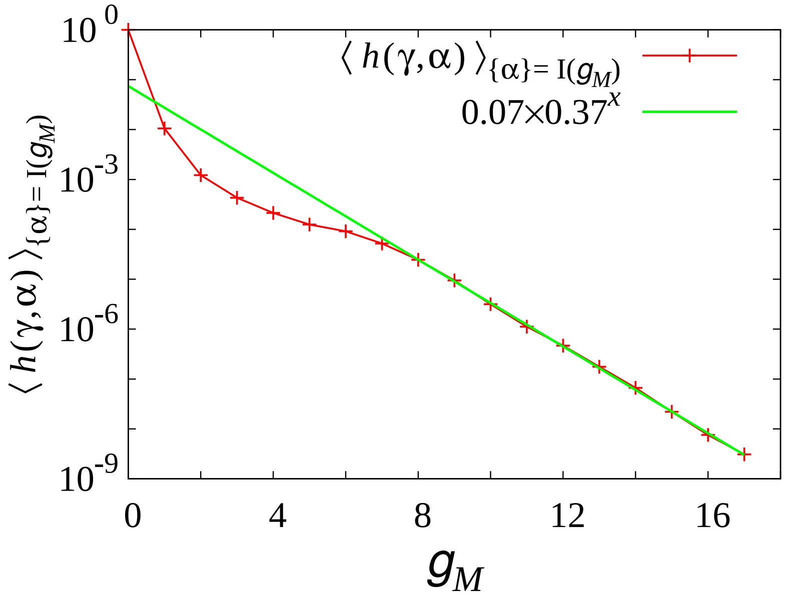


図 4.24: 最短世代における、遺産の平均値の世代依存性。横軸が最短世代、縦軸が遺産の平均値を表す。

ティングしたものを表す。得られた指数は0.37であり、年代世代において求めた値よりも、平均子数などから求めた遺産の平均値の減衰率0.42にわずかではあるがより近い結果が得られた。

次に、年代世代と最短世代における遺産の分布の世代依存性をそれぞれ図 4.25、4.26に示す。 横軸は規格化された遺産の値、縦軸はその確率を表す両対数プロットである。年代世代の $g_Y = 25$ や、最短世代の $g_M = 16$ の分布はそれぞれ別の理由により不十分となるであろうと仮定したが、遺産の平均値が指数関数で上手くフィッティング出来る範囲であったため載せた。図 4.25 の $g_Y = 25$ の分布は、他世代の分布に比べて平均値より値の小さい割合が高くなっている。ただし、図 4.6 から $g_Y = 25$ 以降の個体はまだ増加傾向にあることが分かるため、他世代の分布と異なる特徴を示すのは $g_Y = 25$ での母集団が他世代に比べて不完全であることが原因の一つと考えられる。$g_Y = 25$ 以外の分布では、平均値より値が低い確率はある程度同じ形を示しているように見える。しかし、それに対して平均値より値が高い範囲の確率が世代を追うごとに増加しているため、この分布はまだ収束していないと考えられる。

図 4.26は遺産の分布の最短世代依存性を表す。これら二つの分布は平均値付近の確率が周辺に比べて高いという特徴を有している。この特徴は子数分布が幾何分布の森において確認できており（図 3.16 参照）、また競走馬のメスの子数分布は幾何分布に近い分布を示す（図 4.11 参照）。このため遺産の分布は子数分布による影響を強く受けているのではないかと考えられる。図 4.26での遺産の分布は、年代世代の分布と異なり平均値より低い値の確率が高くなっていく傾向が見られる。最低世代の昇順に馬のデータを取得しているため $g_M = 16$ の個体は全て取り終えている。しかし、図 4.8 から $g_M = 16$ の個体はまだ親子関係を調べ終えていない個体が2割ほど存在するため、データの取得が進み親子関係の情報を新たに得ることによって、分布の最低値や値の低い確率は変わると考えられる。図 4.5 より、$g_M = 15$ にはエラーデータの影響と考えられる個体を無視しても、現代である $g_Y = 29$ （2010〜2013年）に達してしまう。したがって、$g_M = 15$ 前後の世代から将来的には生まれるであろう個体が母集団に含まれておらず、最短世代における遺産の分布には境界条件による効果が含まれていると考えられる。これは現時点での年代世代では起こらない効果であるため、最短世代において世代を追うごとに分布の -4 付近の平均値より値が低い確率が大きくなるという特徴は、境界条件による効果が含まれると

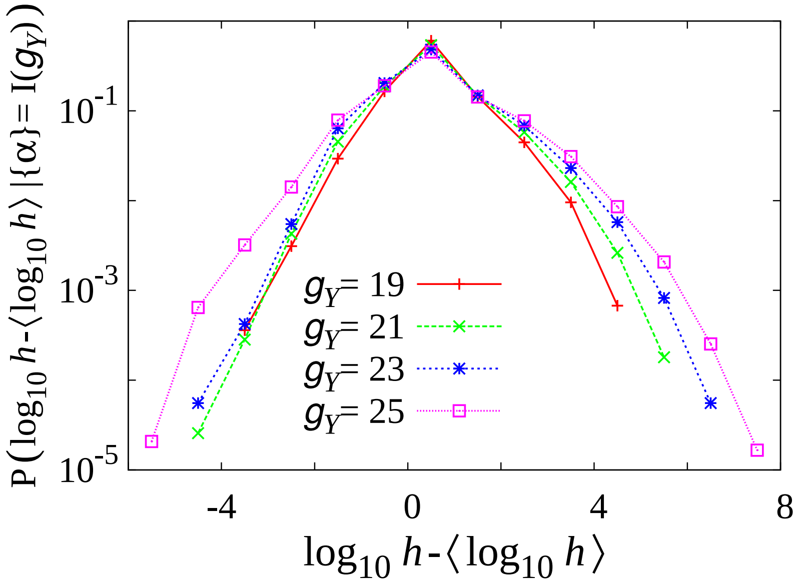


図 4.25: 遺産の分布の年代世代依存性の両対数プロット。横軸が遺産、縦軸が確率を表す。

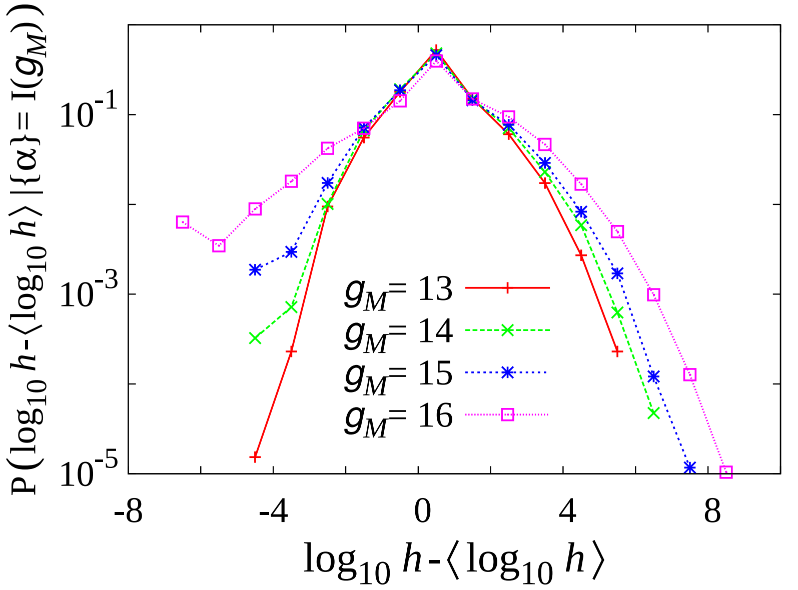


図 4.26: 遺産の分布の最短世代依存性の両対数プロット。横軸が遺産、縦軸が確率を表す。

考えられる。また、最短世代における遺産の分布においても、分布に平均値に比べ高い値の確率はまだ収束していないため、競走馬のデータが増えることによりまた違った特徴が現れるとも考えられる。

第5章　まとめと今後の課題

5.1　まとめ

本研究では、子孫方向に家系図を辿る際に、Derrida らの定義したウェイトに対応する量である関係性と遺産を定義した。それらの量を用いて Derrida らの用いたモデルで子孫方向の木を解析したところ、ウェイトと関係性は、十分世代が離れれば祖先個体には強く依存するが子孫個体には依存しないという特徴が対応した。しかし、遺産はウェイト同様、祖先個体に強く依存するという特徴は一致したが、遺産は子孫個体にも強く依存することが分かった。関係性と遺産の平均値から得られる分布は十分世代が経つことで収束し、その分布の値が微小な領域においてべき則を示すことが分かった。ただし、その指数は異なる値を示した。中立モデルにおいて、子数分布を幾何分布に変更して、遺産の分布について解析を行った。その結果、遺産の分布は子数分布がポアソン分布の場合と同様に、十分世代を経ることである形に収束することが確認できた。また、遺産の平均値の分布において、値が微小な領域においてべき則を示す結果が得られたが、ポアソン分布で得た指数とは一致しなかった。一変数で決まる単純な子数分布の森を構築し、子数分布とべきの指数の関係を調べた。子数が 0 の個体の割合と、平均子数を揃えた分布を用いて遺産の平均値の分布の比較を行った。

競走馬のデータを取得し、年代世代と最短世代という二通りの世代の定義を用いて、競走馬における家系図の構造を解析した。その結果、競走馬は雌雄の数比に差があり、また子数分布がそれぞれ異なること、世代の重複が起きていること、親子間の年齢差が年々小さくなっていることを明らかにした。関係性の分布を求めたところ、正規分布ではない特徴を示した。これは、競走馬の森において配偶がランダムではなく恣意的であることが原因であると思われる。遺産の分布を求めた結果、現時点で取り出せた競走馬の情報では収束していないが、子数分布に依る影響だと

考えられる特徴を示した。

5.2 今後の課題

競走馬の子数分布や数比・親子間の年齢差の減少傾向など、より競走馬の生物集団を再現できるであろう情報を得ることが出来たため、これらの条件をモデルに実装することで、モデルの条件とネットワークの構造との関連性を解析することが考えられる。また、今回の解析の過程で、競走馬の森は閉じた生物集団ではなく、いくらかの流入があることが判明している。これをモデルに実装しその効果を解析することも考えられる。親子間の年齢差の減少傾向が起きている原因の究明と、それに合わせて、年代世代の取り方を改良することにより起こる効果や、年代世代と最短世代の関係も興味深い。関係性の分布の形から、配偶の条件がランダムではないという仮説を立てたが、競走馬に起きている配偶の条件を検討することや、配偶の条件による効果をモデルに実装した解析も今後の課題である。子数分布を変更した生物集団においても、ポアソン分布の場合と同様べき則を観測することは出来たが、関係性や遺産の平均値の分布を特徴づける関数、特に D 分布・D′ 分布の重要な特徴の一つである、値が微小な範囲でべき則を示す条件や、べきの指数を求める関数を見出すことは叶わなかった。実データとモデルの比較と合わせて、これらの特徴を記述する関数についてもさらなる検討が必要である。これらの解析を進めていくことで、ヒトの間に行われる遺産の相続の効果や将来的な予測、生物集団に代わる有向非巡回グラフのネットワーク構造の理解を深めていきたい。

謝辞

本研究を行うにあたり、指導教官である水口毅准教授に心から感謝いたします。水口准教授からは普段の研究活動において多くの助言や指導を賜りました。研究で行き詰った際には研究内容に限らず、精神的にも数多くの場面で支えてくださりました。特に、修士論文の執筆においては数えきれないほどのご助力を賜り、感謝の気持ちが絶えません。大同寛明教授には研究に関わる多くの助言を頂き、また学生生活における問題に対して親身に相談に乗っていただきました。大変感謝しております。また、堀田武彦准教授には多くの場面で本研究に関わるご指摘を頂き、研究活動の糧とさせていただきました。福田浩昭講師には多くの指摘を頂き、考える機会をいただきました。神戸芸術工科大学の大内克哉准教授からは研究や論文紹介など数多くの場面でご指摘を頂きました。先生方の皆様には大変感謝をしております。

右衛門佐誠先輩からは研究内容や研究室生活において多くのご助言をいただきました。大変感謝しております。多くを学ばせてくださった同期の笠間晃氏、西尾一穂氏、研究姿勢に刺激を与えてくださった後輩の皆様への感謝と、こうした学生生活という機会を与えてくださった両親や親族の皆様への感謝の意を結びとさせていただきます。

参考文献

[1] 国連人口基金東京事務所 http://www.unfpa.or.jp/

[2] B. Derrida, S. C. Manrubia and D. H. Zanette, *Phys. Rev. Lett.* **82**, 1987-1990 (1999).

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

[4] 西村麻衣子, 水口毅, 数理解析研究所講究録 **1597**, 191-197 (2008).

[5] 堀内陽介, 水口毅, 守田智, 数理解析研究所講究録 **1663**, 11-13 (2009).

[6] 堀内陽介, 水口毅, 守田智, 数理解析研究所講究録 **1704**, 61-67 (2010).

[7] 堀内陽介, 水口毅, 守田智, 数理解析研究所講究録 **1751**, 85-88 (2011).

[8] 生田成望, 水口毅, 数理解析研究所講究録 **1853**, 56-62 (2013).

[9] 西村麻衣子, 大阪府立大学卒業論文 (2006).

[10] 堀内陽介, 大阪府立大学卒業論文 (2009).

[11] 堀内陽介, 大阪府立大学修士論文 (2011).

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

[13] S. Gualdi, C. H. Yeung and Y. -C. Zhang, *Phys. Rev. E* **84**, 046104 (2011).

[14] M. Matsuo, *Phys. Rev. E* **83**, 066110 (2011).

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