数理解析研究所講究録
第1704巻 2010年 61-67

仮想的両親性生物集団の家系図ネットワークの構造解析

大阪府立大学大学院工学研究科数理工学分野　堀内 陽介
大阪府立大学大学院工学研究科数理工学分野, 科学技術振興機構さきがけ　水口 毅
静岡大学工学部システム工学科　守田 智

1　はじめに

家系図とは生物個体の親子関係を線であらわした図である。一口に家系図と言っても、ある個体の子孫のみに注目した図や、父子のつながりにのみ注目した図など様々なものがあり、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。

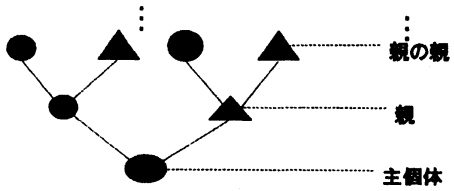

図 1: 単純な家系図の例。●はオス (男) で▲はメス (女) を表す。

本研究の目的は、ある個体の直接の先祖が構成する家系図の構造解析である。まず用語の定義をしよう。有性生殖を行う生物の個体集団において、親子関係にある個体同士を線で結んだものを「森」と呼ぶ。また着目する任意の個体を「主個体」とし、主個体の直接の先祖だけを取り出した図を「木」と呼ぶ。本研究ではこの木の構造に着目する。一般に、木は複雑な構造を持っているが、それは主個体の先祖数に注目するとよく分かる。図1のように有性生殖を行う生物における個体の先祖数は親が2、祖父母が4…と世代を遡る度に2のべき乗で増加していく。つまりG世代前の先祖数を $A(G)$ とすると

$$A(G) = 2^G, \tag{1}$$

となる。

ヒトを例に考えると、30世代前 (1世代を30年と考えるとおよそ900年前) では現代における1個体の先祖は、(1) 式より $2^{30} \approx 10$ 億人存在していたことになる。この見積りは、当時の人口のおよそ3億人という見積りと矛盾している。この矛盾は、実際の家系図は図

2で示されるように、ある先祖が一本の木の中で複数の役割を果たしている、と考えることで解消される。図2を見れば分かるように、複数の役割を持つ先祖が存在すれば、さらにその先祖は必ず役割が重複している。したがって重複した役割を持つ先祖の数は、過去に遡る度に増大していくと考えられ、逆に、正味の先祖個体数は、(1)式の見積もりよりも少なくなることが分かる。このように、一本の木は過去に遡るにつれて複雑に絡み合ったネットワーク構造をなしている。

木の構造に関する先行研究としてDerridaら[1] [2]が、単純な家系図モデルを用いた解析を行っており、西村[3]、堀内[4]、水口ら[5][6]が、それぞれPedigree Online Thoroughbred Database上の競走馬の実データ[7]を用いて、1本の木および2本の木の構造の解析を行っている。西村がG世代前の重複を考えた正味の先祖数に着目し、Derridaの理論的見積もり$A_D(G)$と競走馬の実データ$A_t(G)$との比較を行ったが、その結果、$A_D(G)$と$A_t(G)$が一致していないことが分かった[3][5]。違いの原因は、性差や両親の組み合わせ方など、モデルと競走馬の繁殖過程の間の様々な差異にあると考えられる。本研究では、Derridaらのモデルに「性別の数比を変化させる」という修正を加え、Derridaモデルと競走馬の実データとの違いが性別の数比にあるかどうかの検証を行った。

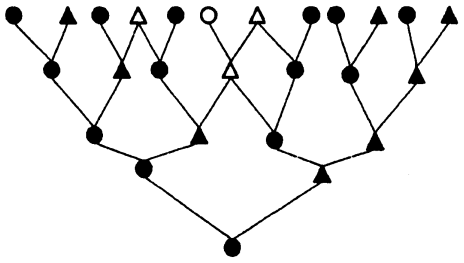

図 2: 複数の役割をもつ個体を含む木の例。●はオス、▲はメスを示し、白抜きで示されたものは役割の重なる個体を示す。

2 家系図について

競走馬の家系図とDerridaモデルの家系図は様々な点で構造が異なっている。2.1節と2.2節でそれぞれを簡単に紹介し、比較する。続いて2.3節で性別の数比の変化を考慮に入れた本研究の家系図モデルを紹介し、2.4節で解析の結果を述べる。

2.1 Derridaモデル

Derridaらは森を構築する過程で、まず最も若い世代を$G=0$とし、G世代前の個体数$N(G)$を

$$N(G) = \left(\frac{2}{m}\right)^G N(0), \tag{2}$$

で計算した。m はペアの平均出産数である。(2) 式にしたがって、世代ごとに個体数を割り振り、個体集団を作る。このとき、オスとメスの数比は 1 : 1 である。次に、各世代内で、オスとメスのペアをランダムに作る。孤立した個体は存在しない。G 世代前の個体の両親を、$G+1$ 世代前のペアからランダムに選ぶ。このようにして作られた親子関係は以下の 4 つの性質を有している。

1. 両親のペアはランダム
2. 性別の数比はオス:メス= 1:1
3. 世代の重複は無い
4. 仔の数は平均 m のポアソン分布に従う

例えば、$N(0) = 10, m = 2$ の時は図 3 のような森が構築される。なお、各世代で親となるペアを作らず、父親母親をそれぞれ $G+1$ 世代のオス・メスからランダムに選ぶという方法でも以下に述べる結果は変わらない。

構築された森において Derrida は、$G = 0$ で任意に主個体を選んだとき、G 世代前の個体数 $N(G)$ のうち主個体の先祖でない割合を非先祖率と呼び、各世代の非先祖率 $s(G)$ が、

$$s(G+1) = \exp[-m + ms(G)], \qquad (3)$$

$$s(0) = 1 - \frac{1}{N(0)}, \qquad (4)$$

で与えられることを理論的に示し、数値的に確かめた。(4) 式は $G = 0$ での非先祖率であり、主個体以外の個体は先祖でないことを意味する。(4) 式を (3) 式に代入すると帰納的に $s(G)$ を求めることができる。例えば $m = 2.5$ の時の (3) 式の結果は図 4 の通りである。図 4 を見ると、$s(G)$ は $G \to \infty$ で s^* に収束すると分かる。$m = 2.5$ の時、$s^* \fallingdotseq 0.1073$ になる。

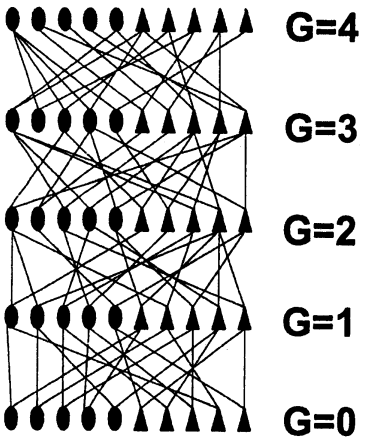

図 3: Derrida モデルによる森。未来方向は下向き。●はオス、▲はメスを示し、親子関係にある個体同士を線で結ぶ。$N(0) = 10, m = 2$。

2.2 競走馬の家系図

競走馬の繁殖過程には人の手が加えられているため、競走馬特有の状況で森が構成されていると考えられる。水口が競走馬の実データから 4349 頭をランダムにサンプリングし、性別の数比と生年分布を見積もった。性別の数比は約 1 : 3 と見積もられている。また、生年分布は時間に対し指数関数で良く近似され、1 年あたりの個体数の増加率は 1.02 である。別に測定した平均出産年齢 12 年を 1 世代とすると、1 世代ごとの平均出産数 $m \fallingdotseq 2.5$

と見積もられている [3][5]。競走馬の実データによる森は、Derrida らのモデルと以下の点で異なり、

1′. 両親はランダムでない
2′. 性別の数比はオス：メス ≒ 1：3
3′. 世代の重複がある
4′. 仔の数はポアソン分布でない

という性質を有すると考えられる。水口は各世代の個体数 $N_t(G)$ に対して、主個体の先祖でない個体数を $N_a(G)$ とし、競走馬の実データによる非先祖率 $s_t(G)$ を

$$s_t(G) = \frac{N_a(G)}{N_t(G)}, \tag{5}$$

で計算した。その結果、$s_t(G)$ は 17 世代前でおよそ 0.9 となった。この値は Derrida 理論による理論値 $s(G)$ のおよそ 8.39 倍であり、Derrida 理論と競走馬のデータは一致しているとは言えない。

2.3 性数比可変モデル

前節で述べた通り、Derrida の単純なモデルでは競走馬のデータを説明できているとは言い難い。この原因は Derrida の仮定 1〜4 のいずれかが破れている事にあると考えられる。そこで本研究においては性別の数比に注目し、前節で挙げた Derrida モデルの仮定の中で、性別の数比を $p : 1-p (0 < p < 1)$ とし、p を変化させて非先祖率の G 依存性を調べた。両親は一つ上の世代のオスから父親、メスから母親をランダムに選び、$N(G)$ の計算は Derrida モデルと同様である。このモデルを「性数比可変モデル」と呼ぼう。このモデルが持つ性質をまとめると、

1. 両親のペアはランダム
2″. 性別の数比はオス:メス＝ $p : 1-p$
3. 世代の重複は無い
4. 仔の数は平均 m のポアソン分布に従う

である。

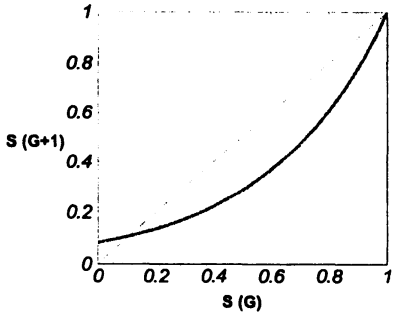

図 4: $s(G)$ の漸化式。$m = 2.5$。横軸は $s(G)$、縦軸は $s(G+1)$。点線は $s(G+1) = s(G)$、実線は $s(G+1) = \exp[-m + ms(G)]$ を示す。この 2 関数の交点が s^*。

2.4 解析

性数比可変モデルにおいて、$N(0) = 5000$、平均出産数mは競走馬の実データをサンプリングすることにより算出した$m = 2.5$を用い、pを0.5(Derridaモデル)、0.25(競走馬の性別の数比)、0.1、0.01、0.001と変化させ、非先祖率$s_r(G,p)$を測定した。それぞれのpに対して10個のサンプルを取り、(5)式と同様にして測定した非先祖率のサンプルの平均$\bar{s_r}(G,p)$を算出した(図5)。同じ森では選択する主個体が違っていても、得られる結果はほとんど変わらない。これは、木の作成をどの主個体から始めても、過去に遡るにつれて木が同じになる傾向があるからだと考えられる[4]。図5には比較のため、競走馬の実データから算出された$s_t(G)$およびDerridaモデルによる$s(G)$も描いた。競走馬の実データから見積もられた現代の個体数$N(0)$は230000であったので、$N(0) = 230000$、$m = 2.5$の時の$s(G)$もプロットしている。$N(0)$の値の変化で$s(G)$の変化がどう変わるのかを見るために、$N(0) = 23000$、$m = 2.5$の時の$s(G)$も計算し、$\bar{s_r}(G,p)$や$s_t(G)$と比較した。

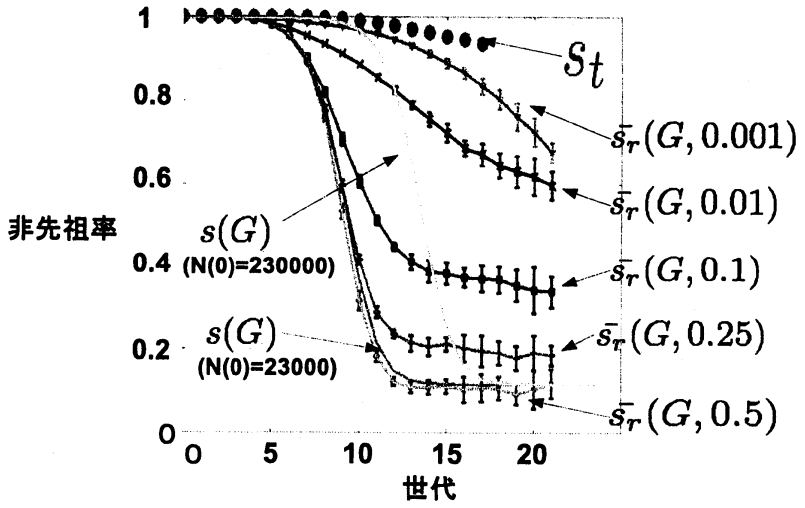

図5: 性数比可変モデルによる、pを変化させたときの$\bar{s_r}(G,p)$のG依存性、競走馬の実データによる$s_t(G)$の実測値、Derridaモデルによる$s(G)$の理論値を表した図。縦軸はそれぞれの非先祖率s, s_t, $\bar{s_r}$、横軸は世代を表す。丸印は$s_t(G)$。それ以外の印は$\bar{s_r}(G,0.001)$、$\bar{s_r}(G,0.01)$、$\bar{s_r}(G,0.1)$、$\bar{s_r}(G,0.25)$(競走馬の性別比に相当)、$\bar{s_r}(G,0.5)$(Derridaモデルに相当)を示す。縦線はそれぞれの値の標準偏差を示す。$N(0) = 23000$、$N(0) = 230000$の時の$s(G)$も示す。

まず、$s(G)$においては、$N(0) = 230000$、$N(0) = 23000$の時の$s(G)$を比較すれば分かるように、$N(0)$が大きいほど値の落ち込みが遅くなるが、収束する$s(G)$の値は変わらな

いことが分かる。次に、図5での性数比可変モデルによる $\bar{s_r}(G,p)$ の値を見ると、p の値が小さくなるほど世代を遡るにつれて起こる $s_r(\bar{G},p)$ の値の落ち込みが小さくなり、収束する値が大きくなることが分かる。そして、$\bar{s_r}(G,0.25)$ と $s_t(G)$ を比べると、$s_t(G)$ の落ち込みは $\bar{s_r}(G,0.25)$ に比べて小さく、また G が大きい時、$s_t(G)$ は $\bar{s_r}(G,0.25)$ の4倍ほどの値に収束しており、二つの値は一致していない。これは、Derridaモデルを性の数比について改良するだけでは、競走馬のデータを説明できないことを意味する。

3 考察

本研究では、性数比可変モデルを用いて、p を変化させたときの非先祖率 $\bar{s_r}(G,p)$ の G 依存性を解析し、Derridaモデルによる非先祖率 $s(G)$ と競走馬の実データによる $s_t(G)$ の実測値の違いが、性別の数比によるものかどうかを調べた。その結果、オスとメスの数比が極端になるほど、つまり p が小さくなるほど、世代を遡るにつれての非先祖率の落ち込みが小さくなり、収束値が高くなることが分かった。また、Derridaモデルと競走馬のデータが異なる原因が、性別の数比だけではないことが示唆された。

ここで本論文で用いた平均出産数 m と、性の数比 p の値の妥当性について考えてみたい。まず平均出産数に関して言えば、本論文では $m = 2.5$ と設定した。これは、ランダムにサンプリングされた4349頭(オス1256頭、メス3093頭)の誕生年分布から見積もった登録数の増加率を元に算出されたものである[5]。一方、同じ4349頭の仔数の直接の平均は、オス30.8頭、メス3.6頭であった。これらの値に性数比の重みを考慮しても $m = 2.5$ にはならない。これはデータベースに登録されなかった馬や、仔数0の個体データの抹消などデータベースのもれが原因であると考えられ、これらを考慮しながら m を設定する必要があると考えられる。同様に性の数比についても、文献[5]での見積もりにより、$p = 0.25$(オス：メス=1：3)と設定したが、日本軽種馬登録協会[8]に登録されている産駒160023頭(1982年～2008年生まれ)のうちオスは80781頭であり、ほぼ数比は1：1である。こちらの原因もデータベースのもれ等が考えられ、$p = 0.5$ とした方がいいのかもしれない。

未解決な点の一つに、$G \to \infty$ の時の $\bar{s_r}(G,p)$ の収束値の p 依存性があり、その解明に現在取り組んでいる。そのためには、収束したとみなすことができるまでさらに昔の世代に遡り解析する必要がある。

最後に、本研究で用いるモデルの拡張について考える。本論文で解析した性数比可変モデルの結果は、競走馬のデータと合っていなかった。これは競走馬の繁殖過程が複雑であり、Derridaモデルの性数比に関する改良だけでは、うまくモデル化できなかったからだと思われる。競走馬の家系図の構造を明らかにするために、他の仮定(1,3,4)も変え、より複雑な条件を考慮したモデルを用いて解析することも考えられる。

謝辞

我々の研究に対し、惜しみない助言と提案を行っていただいた大同寛明教授、堀田武彦准教授、福田浩昭助教、私の研究室の学生の方々に感謝しております。

参考文献

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

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

[3] 西村麻衣子, 大阪府立大学卒業論文 (2006).

[4] 堀内陽介, 大阪府立大学卒業論文 (2009).

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

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

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

[8] 日本軽種馬登録協会 http://www.studbook.jp/ja/.