数理解析研究所講究録
第1751巻 2011年 85-88

家系図ネットワークに対する性比と仔数分布の効果

大阪府立大学大学院工学研究科数理工学分野数理工学分野 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 問題設定

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. この中で, ある個体に注目した時, その祖先数は世代をさかのぼるごとに2のべき乗で増加する. こういった祖先数の指数関数的の増加は, 「祖先数が過去のある世代の総個体数を超える」というパラドクスを生じる. このパラドクスは, 「先祖個体の中に複数の役割を持つ個体が存在する」と考えることで解消されるが, この事実により家系図の構造が複雑になっていると考えられる. 本論文ではこの家系図の構造を研究対象にする. 具体的には, 着目する任意の個体を「主個体」, 主個体の直接の祖先だけを取り出したものを「木」と呼び, 家系図の中でも主個体の祖先に注目し, 一本の木のネットワーク構造を解析した.

この構造に関する研究がいくつか報告されている. Derrida ら [1][2] は, 非先祖率やウェイトといった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. Pla ら [3] は Derrida らのモデルを一般化したモデルを用いている. 水口 [4][5], 西村 [6] や筆者 [7][8][9] が実データの家系図における木の構造を解析している. そして西村や筆者, 水口は実データとして競走馬の家系図を用いてその非先祖率を実測したが, Derrida らの解析結果と実データを用いた解析結果は異なっていることが分かった. 違いの原因は, 性差や両親の組み合わせ方や設定する仔数分布の違いなど, モデルと競走馬の繁殖過程の間のさまざまな差異にあると考えられる.

本研究では非先祖率を解析することで家系図の構造を見ていく. 加えて, Derrida らのモデルに「性の数比及びオスとメスの仔数分布を変化させる」という修正を加えることで, Derrida らのモデルと競走馬の家系図をつなぐようなモデルを作ることができないかを考え, その条件を考慮したモデルを設定して, 仮想的な生物集団の家系図—以下森と呼ぶ—を構築し, 非先祖率を解析した.

2 性比-仔数分布可変モデルと解析

先行研究や本研究で用いられた G 世代前の非先祖率や世代個体数を区別したものを表1にまとめておく. Derrida らのモデル及び解析は文献 [1][2], Pla らのモデル及び解析は文献 [3], 競走馬データの解析及び Derrida モデルと競走馬の家系図の比較は文献 [4 - 9] を参照されたい. Derrida らのモデルと競走馬の家系図をつなぐために, まず我々は Derrida らのモデルを性の数比を変化できるように改良した. このモデルを「性数比可変モデル」と呼び非先祖率を解析したが, このモデルでは Derrida らのモデルと競走馬データの違いを説明できなかった [7][9].

そこで我々は性の数比に加え仔数分布にも注目した. すなわちオスが k 頭の仔を持つ確率を $P_m(k)$, メスが k 頭の仔を持つ確率を $P_f(k)$ とし, それぞれにいろいろな仔数分布を与えて森を構築し, 非先祖率 $s_h(G)$ の世代依存性を調べた. 計算機上で森を構築する過程で, Derrida らのモデルと同様に, 最も若い世代を $G=0$ とし, 世代の重複はないとした. 性比 p をオスが産まれる確率とし, 世代を通して性比 p は固定する. Derrida らのように, 過去に遡る方向で森を作る時に 両親を1つ上の世代からランダムに選べば, その結果各個体の仔数分布はポアソン分布になる. しかし, 過去に遡る方向に森を構築する方法で両親の選び方に工夫をし, 与えられた仔数分布を作るのは難しい. そこで Derrida モデルや性数比可変モデルとは異なり最も昔の世代の総個体数を与え, $G=0$ に向かって未来方向に森を作る. 世代ごとにオスの仔数とメスの仔数を, それぞれ $P_m(k)$, $P_f(k)$ に従って個体ごとに割り振る.

表 1: 各モデルやデータで使われる G 世代前の非先祖率と世代個体数

モデル	非先祖率	世代個体数
Derrida ら	$s(G)$	$N(G)$
競走馬データ	$s_T(G)$	$N_T(G)$
性比可変	$s_f(G)$	$N_f(G)$
性比-仔数分布可変	$s_h(G)$	$N_h(G)$
非先祖率の理論値	$s'(G)$	—

1 つ下の世代の個体数は, オスの仔数の合計とメスの仔数の合計が一致しない時は合計数が少ない性別の総仔数が採用される. その時採用されなかった性別の仔数分布は, 採用した個体数を総仔数として各個体に割り振られる. 両親は、1 つ上の世代のオスから一様乱数 [10] により父親, メスから母親をランダムに選んだ. 本研究において作られる家系図は以下の特徴を有している.

1 . 両親のペアはランダム

2′ . 性別の数比はオス:メス=$p : 1-p$

3 . 世代の重複は無い

4′ . 仔の数はオスは $P_m(k)$、メスは $P_f(k)$(k:仔数) に従う

このモデルを「性比-仔数分布可変モデル」と名づけた. ′ 印は Derrida らのモデルとの相違点を表している.

解析に使用した $P_m(k)$, $P_f(k)$ の組み合わせを表 2 に示す. 各組み合わせごとで森を 10 個作り, その平均を計算した. なお, 「カットオフ-べき分布」は筆者らが競走馬のデータベース上 [11] から競走馬のデータをサンプリングし, $P_m(k)$ をフィッティングし,

$$P_m(k) = \begin{cases} a_0 & (k=0), \\ a_1 k^{-\gamma} & (0 < k \leq k^*), \\ 0 & (k > k^*), \end{cases} \tag{1}$$

と見積もったものである. 式 (1) 内のパラメーターはそれぞれ,

$$a_0 = 0.9, \quad a_1 = \frac{1-a_0}{\sum_{k=1}^{k=k^*} k^{-\gamma}}, \quad \gamma = 1.2, \quad k^* = 1000, \tag{2}$$

である. メスの方は競走馬データのサンプリングにより得られた $P_f(k)$ の見積もりから得られたものを,

$$P_f(k) = (1-q)q^k, \tag{3}$$

とフィッティングし, この幾何分布を用いている. パラメーター $q = 0.71$ である. ポアソン分布のパラメーターは, 競走馬データの個体増加率の見積りから $m = 2.5$ とした. また競走馬の性比 p を, $p = 0.4$ と見積もった. 以上の競走馬データの解析の詳しい記述は文献 [9][12] を参照されたい. 競走馬の実データから $N_T(20) \sim 2600$ と見積もられたが, 性比-仔数分布可変モデルにおいては解析の便宜上 $N_h(20) = 200$ とした.

非先祖率 $s_h(G)$ の平均値 $\bar{s}_h(G)$ の世代依存性及び競走馬の実データによる $s_T(G)$ の実測値を図 1 に示す. 図 1 を見ると, 仔数分布によって非先祖率の振る舞いが大きく変化していることがわかる. とくにカットオフ-べき分布の影響は大きく, 組み合わせ 1 や組み合わせ 2 のようにオスの仔数分布にカットオフ-べき分布を導入した時は $\bar{s}_h(G)$ の値がかなり高くなる. モデルの性比と仔数分布を実データのものにあわせることで, 非先祖率の世代依存性が実データの振る舞いにかなり近くなることが分かった. 具体的には組み合わせ 1 の時のオス・メスの仔数分布を採用した時の $\bar{s_h}$ の 17 世代前の値はおよそ 0.769 となったが, この値は $s_T(G)$ のおよそ 0.854 倍であり,

表 2: $P_m(k)$ と $P_f(k)$ の組み合わせ

組み合わせ	$P_m(k)$	$P_f(k)$
1	カットオフ-べき分布	幾何分布
2	カットオフ-べき分布	ポアソン分布
3	ポアソン分布	幾何分布
4	ポアソン分布	ポアソン分布

Derrida による見積り $s(G)$ と比較すると競走馬の実データの振る舞いに近くなっている.

また, 性比-仔数分布可変モデルにおいても, 非先祖率を理論的に見積もることができる. G 世代の非先祖率を $s'(G)$ とすると,

$$s'(G+1) = p\phi_m(s'(G)) + (1-p)\phi_f(s'(G)), \quad (4)$$

が成り立つ. $\phi_m(x)$, $\phi_f(x)$ はそれぞれ $P_m(k)$, $P_f(k)$ の確率母関数である. 式 (4) を使った非先祖率の理論値 s' を図 2 に示す. 競走馬のデータから見積もった $G=0$ での総個体数 $N_T(0) = 230000$ を用いて, $s'(0) = 1 - \dfrac{1}{230000}$ とした. 図 2 を見ると, 理論的見積もりにおいてもカットオフ-べき分布の影響は大きく, これを用いた s' の値が大きくなり, 組み合わせ 1 の時の非先祖率の収束値は実データに最も近くなっている事が分かる.

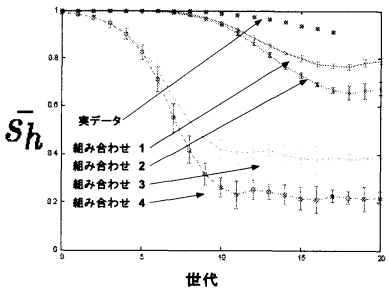

図 1: 性比-仔数分布可変モデルによる、オスとメスの仔数分布を変化させた時の $\bar{s}_h(G)$ の世代依存性, 競走馬の実データによる $s_T(G)$ の実測値を表した図. 縦軸はそれぞれの非先祖率 $\bar{s_h}$, $s_T(G)$, 横軸は世代を表す. 印はそれぞれ $s_T(G)$, 表 2 で示した仔数分布の組み合わせの時の非先祖率の世代依存性を表している. 各印の縦線は標準偏差を示している.

3 まとめと今後の課題

本研究では様々な条件を考慮してモデルを構築して森及び木を構成し, 非先祖率を解析した. 各個体に与える仔数分布の関数を変化させながら非先祖率を計算したが, 仔数分布の違いは非先祖率の世代依存性に大きな変化を与えることが分かった. 仔数分布は家系図ネットワーク構造に大きな影響を及ぼしていると言える. また, 競走馬の実データからフィッティングした性比や性別ごとの仔数分布などのパラメーターを本研究の家系図モデルに適用させ, 非先祖率のシミュレーション値と競走馬データによる非先祖率の実測値を比較した, Derrida らのモデルや性数比可変モデルによる非先祖率の解析と比べると、性比-仔数分布可変モデルによる非先祖率の世代依存性はかなり実データのそれに近づいており, 実データの家系図構造を再現できたと考えられる.

本研究では競走馬の家系図構造と, Derrida らのモデルによる家系図をつなぐようなモデルを構築することを目的の 1 つとしたが, 競走馬の実データから見積もられたパラメーターの設定にはさらなる解析が必要である. バイアスのかかっていないデータのサンプリングを行ったり, 競走馬を実際に扱う牧場などに聞き込みやフィールドワークによって, より偏りの無いデータを取ることが必要である. できるだけ正確なパラメータを用いてモデルの再構築を行い, 競走馬の家系図構造に近づけることが重要だと考えられる.

また性比-仔数分布可変モデルにおいて, ある世代でのオスの持つ仔数の合計とメスの持つ仔数の合計が一致しない時は少ないほうが 1 つ下の世代個体数として採用されるようにしているが, それにより採用されなかった性別の仔数分布が設定時の仔数分布からずれてしまっていると考えられる, シミュレーションによる誤差をより少なくできるモデルの改良を考えることも重要であると考えられる. さらにこのモデルでは, Derrida ら, Pla らのモデルと同様に世代の重複は無いという仮定のもと森を構築しているが, 競走馬の家系図には世代の重複ももちろん含まれている.

世代の重複を考慮した森の構築を考えるのも重要であると考えられる.

また本研究では非先祖率の解析を通して森の構造を見てきた. そして本研究を通して非先祖率の振る舞いの変化は性比や仔数分布などのパラメーターの値の変化によって引き起こされていることが分かった. ではその非先祖率の振る舞いの変化の要因となるこれらのパラメーターの値は. 何によって決まっているのであろうか. 競走馬の場合であればレースの戦績や, 各個体を引き取る際の取引額の高低, 生物的な特性等様々な要因が絡み合って仔数分布や性比などのパラメーターが決まると考えられる. これらのパラメーターを決める要因をモデル設計に取り込んで森を作り, 非先祖率の解析を行えるようにしたい.

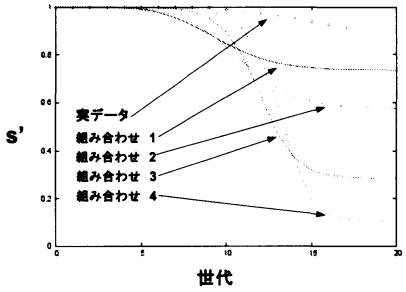

図 2: 式 (4) による非先祖率の理論的見積り. 縦軸は s', 横軸は世代を表す. 印は $s_T(G)$. それ以外の実線は表 2 で示した仔数分布の組み合わせの時の s' の世代依存性を表している.

謝辞

我々の研究に対し, 惜しみない助言と提案を与えてくださった大同寛明教授, 堀田武彦准教授, 福田浩昭助教, 研究室の学生の方々に感謝の意を表します.

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**, 303-305.

[3] P. De L. Rios and O. Pla, Phys. Rev. E. (2000) **61**, 5620-5623.

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

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

[6] 西村麻衣子, 大阪府立大学卒業論文 (2006).

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

[8] 堀内陽介, 大阪府立大学卒業論文 (2009).

[9] 堀内陽介, 大阪府立大学大学院修士論文 (2011).

[10] William H. Press, William T. Vetterlinq, Saul A. Teukoisky, Brian P. Flannery, *NUMERICAL RECIPES in C* /日本語版/, 技術評論社 (1993).

[11] Pedigree Online Thoroughbred Databese http://www.pedigreequery.com/.

[12] S.Morita and T.Mizuguchi, private communication.