Wright–Fisher Model

岡田崇

2026 年 6 月 14 日

目次

本ノートの目的：集団遺伝学の基本的な問い

集団遺伝学（population genetics）は，集団の中にある遺伝的なタイプの割合が，世代を通じてどのように変わるかを調べる分野である．たとえば，集団の中に二つのタイプ A と a があるとして，タイプ A の個体数や頻度が，時間とともに増えるのか，減るのか，あるいは偶然に消えてしまうのかを考える．

このノートでは，そのためのもっとも基本的なモデルの一つである Wright–Fisher model を導入する．出発点は，「次の世代の個体は，現在の集団からランダムに親を選んで作られる」という単純な考え方である．集団サイズが有限である限り，たとえタイプ間に有利・不利がなくても，誰が次世代に子孫を残すかには偶然が入る．この偶然が積み重なると，タイプの頻度は揺らぎ，ときには一方のタイプが完全に消えたり，逆に集団全体に固定したりする．

このような有限集団におけるランダムな頻度変化を **genetic drift**（遺伝的浮動）という．一方で，タイプごとに増えやすさが異なると，頻度変化には平均的な方向づけが生じる．これが **selection** である．また，あるタイプから別のタイプへ変わることを **mutation** という．集団遺伝学では，genetic drift，selection，mutation が組み合わさることで，集団の遺伝的組成がどのように変わるかを考える．

本ノートでは，ウイルスや微生物進化を念頭に，**haploid** 集団に限定する．したがって，各個体は一つの遺伝子コピーだけを持つと考える．特に，まずは二つのタイプ A と a だけを持つ **bi-allelic model** を基本とする．

1 基本設定

集団サイズを一定の N とする．時刻 t におけるタイプ A の個体数を

$$I_t \in \{0, 1, \dots, N\}$$

と書く．タイプ A の頻度を

$$X_t = \frac{I_t}{N}$$

とする．タイプ a の頻度は $1 - X_t$ である．

記号の読み方

I_t は「個体数」であり，X_t は「割合」である．シミュレーションでは個体数 I_t を更新するが，理論解析では割合 X_t を見ることが多い．なぜなら，X_t は常に 0 から 1 の範囲に入り，集団サイズの違う状況も比較しやすいからである．特に $X_t = 0$ はタイプ A が消失した状態，$X_t = 1$ はタイプ A が固定した状態を表す．

以降では，タイプ A の selection coefficient を s とし，fitness を

$$w_A = 1 + s, \qquad w_a = 1$$

とする．また，mutation rate を

$$A \to a : u, \qquad a \to A : v$$

と書く．

2 Moran model と Wright–Fisher model

Moran model と Wright–Fisher model は，どちらも有限集団における genetic drift を記述する代表的な stochastic model である．ただし，両者では「1 step が何を意味するか」が異なる．そこでこの節では，まず中立な Wright–Fisher model と中立な Moran model をそれぞれ定義し，その後で更新単位と時間スケールの違いを整理する．

2.1 中立な Wright–Fisher model

中立（neutral）とは，タイプ A とタイプ a の間に fitness 差がないことをいう．すなわち，

$$w_A = w_a = 1$$

である．

Wright–Fisher model では，各世代で次世代の N 個体を，現在世代から独立に復元抽出する．現在のタイプ A の個体数が $I_t = i$，頻度が $x = i/N$ のとき，次世代のタイプ A の個体数 I_{t+1} は

$$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, x = i/N)$$

に従う．ここで記号 $\sim$ は “is distributed as” を意味する．したがって，$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, x)$ は，条件 $I_t = i$ のもとで，次世代のタイプ A 個体数 I_{t+1} が parameters N, x の binomial distribution に従う，という意味である．一般に $Y \sim \mathrm{Bin}(n, p)$ は，成功確率 p の独立な試行を n 回行ったときの

成功回数 Y の分布を表し，

$$\mathbb{P}(Y=k)=\binom{n}{k}p^k(1-p)^{n-k},\qquad k=0,1,\dots,n$$

である．Wright–Fisher model では，N 個の子個体を独立に作り，それぞれがタイプ A になる確率が x である，という sampling を表している．したがって transition probability は

$$\mathbb{P}(I_{t+1}=j\mid I_t=i)=\binom{N}{j}\left(\frac{i}{N}\right)^j\left(1-\frac{i}{N}\right)^{N-j}$$

である．

conditional mean と variance は

$$\mathbb{E}[X_{t+1}\mid X_t=x]=x,$$

$$\mathrm{Var}(X_{t+1}\mid X_t=x)=\frac{x(1-x)}{N}$$

である．neutral な場合，平均的には頻度は変化しないが，有限集団であるために variance を持つ．このランダムな頻度変化が **genetic drift** である．

Remark: “drift” という用語

本ノートでは，単に drift と書くときは，原則として **genetic drift** を指す．つまり，有限集団における random sampling によって生じるランダムな頻度変化のことである．selection による平均的な方向づけや，mutation による流入・流出とは区別して読む．一方で，物理学の文脈では，決定論的な力や平均的な流れを “drift” と呼ぶことがある．混同を避けるため，以下では selection や mutation が作る項は，できるだけ deterministic term/bias と呼ぶ．

中立モデルは「何も起こらないモデル」ではない

中立では，期待値だけを見ると $\mathbb{E}[X_{t+1}\mid X_t=x]=x$ である．しかし実際の sample path では，binomial sampling によって頻度が毎世代揺らぐ．この揺らぎが積み重なることで，中立なタイプでも消失したり固定したりする．

2.2 中立な Moran model

中立な Moran model では，1 event で親個体と死亡個体を一様ランダムに選ぶ．親個体は同じタイプの子を 1 個体残し，死亡個体は集団から取り除かれる．集団サイズは常に N に保たれる．

現在 $X=x$ のとき，タイプ A が 1 個体増えるのは，親がタイプ A で死亡個体がタイプ a のときである．したがって

$$p^+(x)=\mathbb{P}\left(\Delta X=\frac{1}{N}\mid X=x\right)=x(1-x)$$

である．同様に，タイプ A が 1 個体減るのは，親がタイプ a で死亡個体がタイプ A のときであり，

$$p^-(x) = \mathbb{P}\left(\Delta X = -\frac{1}{N} \mid X = x\right) = x(1-x)$$

である．親と死亡個体が同じタイプの場合，タイプ A の個体数は変化しない．よって，1 event あたり

$$\mathbb{E}[\Delta X \mid X = x] = 0, \qquad \mathrm{Var}(\Delta X \mid X = x) = \frac{2x(1-x)}{N^2}$$

となる．

ここで重要なのは，Moran model では 1 event で頻度が高々 $1/N$ しか変わらないという点である．一方，Wright–Fisher model では 1 世代で N 個体をまとめて sampling するため，1 step あたりの揺らぎの大きさが異なる．

2.3 更新単位と時間スケールの違い

Wright–Fisher model では，1 step で集団全体が次世代に置き換わる．一方，Moran model では，1 event で 1 個体が出生し，1 個体が死亡する．したがって，同じ「1 step」という言葉を使っても，両者では進む時間の量が異なる．

	Wright–Fisher model	Moran model
更新単位	1 世代	1 birth–death event
1 回の更新	N 個体をまとめてサンプリング	高々 1 個体だけが入れ替わる
頻度の変化量	$O(1)$ 個体分が同時に変わりうる	$\pm 1/N$ または 0
自然な用途	世代が明確な集団	世代が重なる集団・連続的な出生死亡

Moran model の 1 event はとても小さい更新であり，Wright–Fisher model の 1 世代とはそのまま比較できない．たとえば Moran model を数十 event だけ回しても，大きな集団では頻度はほとんど変わらない．これは Moran model の genetic drift が弱いという意味ではなく，単に 1 event で進む時間が短いという意味である．

Wright–Fisher model の 1 世代あたりの variance は

$$\frac{x(1-x)}{N}$$

である．一方，Moran model の k event 分の variance は，粗く見れば

$$k \cdot \frac{2x(1-x)}{N^2}$$

である．これらを合わせるには，おおよそ

$$k = \frac{N}{2}$$

とすればよい．つまり，この規約では，**中立な Moran model の $N/2$ event 程度が Wright–Fisher model の 1 世代程度の genetic drift に対応する**．

より標準的には，diffusion limit で次のように書ける．Wright–Fisher model では，世代数を t_{WF} として

$$\tau = \frac{t_{\mathrm{WF}}}{N}$$

とおくと，generator は

$$\mathcal{L}f(x) = \frac{1}{2}x(1-x)f''(x)$$

に収束する．Moran model では，event 数を t_{M} として

$$\tau = \frac{2t_{\mathrm{M}}}{N^2}$$

とおくと，同じ generator

$$\mathcal{L}f(x) = \frac{1}{2}x(1-x)f''(x)$$

に収束する．したがって，どちらも中立な diffusion limit では次の同じ式で近似される．

> **neutral Wright–Fisher diffusion**
>
> $$\mathrm{d}X_\tau = \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

Rescale すると中立 Moran と Wright–Fisher は同じになる

中立な場合，Moran model と Wright–Fisher model の違いは，主に更新単位の違いである．Moran model では 1 event が小さいので，event 数を $2t_{\mathrm{M}}/N^2$ の時間に直す．Wright–Fisher model では世代数を t_{WF}/N の時間に直す．この rescale 後，どちらも同じ Wright–Fisher diffusion

$$\mathrm{d}X_\tau = \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

を与える．

なお，文献によっては「Moran model の 1 世代」を N event と定義したり，連続時間の event rate を変えたりする．その場合，上の対応には定数倍の違いが出る．重要なのは，**Moran model を Wright–Fisher model と比較するときは，event 数そのものではなく，rescale された時間で比較する**という点である．

2.4 このノートでの Moran model の扱い

Moran model は，Wright–Fisher model と同じ genetic drift を別の更新単位で記述するモデルとして重要である．ただし，本ノートの主目的は Wright–Fisher model を理解することである．そのため，Moran model については，中立な場合に時間を rescale すると Wright–Fisher diffusion と同じ極限を持つことを確認するにとどめる．

以降の selection，mutation，diffusion approximation，fixation probability，実装は，基本的に Wright–Fisher model に焦点を当てて説明する．ただし，fixation probability を branching process で直感的に導く箇所では，連続時間 birth–death process や Moran model の初期段階に近い考え方を補助的に使う．

2.5 selection がある Wright–Fisher model

タイプ A の fitness を $1+s$，タイプ a の fitness を 1 とする．現在のタイプ A の頻度が x のとき，selection を受けた後に親として選ばれるタイプ A の確率は

$$q_s(x) = \frac{(1+s)x}{(1+s)x + (1-x)} = \frac{(1+s)x}{1+sx}$$

である．したがって，次世代の A 個体数は

$$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, q_s(x))$$

に従う．

このとき

$$\mathbb{E}[X_{t+1} \mid X_t = x] = q_s(x)$$

である．$s > 0$ ならばタイプ A は有利，$s < 0$ ならば不利である．

selection の読み方

$q_s(x)$ は，単なる頻度 x ではなく，「fitness で重みづけした親の選ばれやすさ」である．$s > 0$ のときは $q_s(x) > x$ となり，タイプ A は次世代の親として少し選ばれやすくなる．ただし，次世代の個体数はなお二項分布からサンプリングされるため，有利なタイプが必ず増えるわけではない．有限集団では，selection による方向づけと genetic drift による偶然が同時に働く．

2.6 mutation がある Wright–Fisher model

mutation rate を

$$A \to a : u, \qquad a \to A : v$$

とする．中立な場合，現在の頻度が x なら，次世代の個体がタイプ A になる確率は

$$p(x) = (1-u)x + v(1-x)$$

である．よって

$$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, p(x))$$

である．

selection と mutation を両方入れる場合，まず selection によって親のタイプ分布が

$$q_s(x) = \frac{(1+s)x}{1+sx}$$

になり，その後に mutation が起こると考えると，次世代個体がタイプ A である確率は

$$p_{s,u,v}(x) = (1-u)q_s(x) + v\{1-q_s(x)\}$$

である．したがって

$$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, p_{s,u,v}(x))$$

となる．

mutation の役割

mutation は，タイプ間の一方向または双方向の流れを作る．たとえば $u > 0$ は A から a への流出，$v > 0$ は a から A への流入である．mutation がない場合，$I_t = 0$ や $I_t = N$ に到達するとそこから抜け出せない．一方，双方向 mutation があると，消えたタイプも再び生じうるため，境界は absorbing state ではなくなる．この違いは，長時間のシミュレーションや stationary distribution を考えるときに特に重要である．

3 Wright–Fisher model の diffusion approximation

3.1 diffusion approximation の考え方

Wright–Fisher model は離散時間・有限状態の Markov chain である．しかし，集団サイズ N が大きいとき，頻度 $X_t = I_t/N$ は $[0,1]$ 上の連続的な stochastic process で近似できる．これを **diffusion approximation** という．

なぜ diffusion approximation を使うのか

有限な N に対する Wright–Fisher model は，状態 $0, 1, \ldots, N$ を持つ Markov chain である．N が小さい場合はこのまま扱えるが，N が大きいと状態数が増え，fixation probability や密度の時間発展を直接計算するのが重くなる．diffusion approximation を使うと，離散的な個体数のモデルを，連続変数 $X_\tau \in [0,1]$ の stochastic differential equation（SDE）として扱える．これにより，fixation probability，mean fixation time，stationary distribution などを解析しやすくなる．

1 世代あたりの頻度変化を

$$\Delta X_t = X_{t+1} - X_t$$

とする．diffusion approximation では，

$$\mathbb{E}[\Delta X_t \mid X_t = x] = \frac{1}{N} b(x) + o\left(\frac{1}{N}\right),$$

$$\mathrm{Var}(\Delta X_t \mid X_t = x) = \frac{1}{N} a(x) + o\left(\frac{1}{N}\right)$$

となるような状況を考える．時間を

$$\tau = \frac{t}{N}$$

とスケールすると，X_t は次の SDE で近似される：

$$\mathrm{d}X_\tau = b(X_\tau)\,\mathrm{d}\tau + \sqrt{a(X_\tau)}\,\mathrm{d}W_\tau.$$

ここで W_τ は標準 Brownian motion である．

この SDE は，個々の sample path を直接記述する形である．一方，同じ diffusion process を「確率密度の時間発展」として見ると Fokker–Planck equation になる．対応する generator は

$$\mathcal{L} f(x) = b(x) f'(x) + \frac{1}{2} a(x) f''(x)$$

であり，密度 $\rho(x,\tau)$ が存在する場合，Fokker–Planck equation は

$$\frac{\partial \rho}{\partial \tau} = -\frac{\partial}{\partial x}\{b(x)\rho\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{a(x)\rho\}$$

である．つまり，SDE form と Fokker–Planck form は別々のモデルではなく，同じ generator に対応する等価な記述である．前者は trajectory を見たいとき，後者は density や stationary distribution を見たいときに便利である．

3.2 中立な bi-allelic Wright–Fisher model からの導出

中立な場合，

$$I_{t+1} \mid I_t = i \sim \mathrm{Bin}(N, x), \qquad x = \frac{i}{N}$$

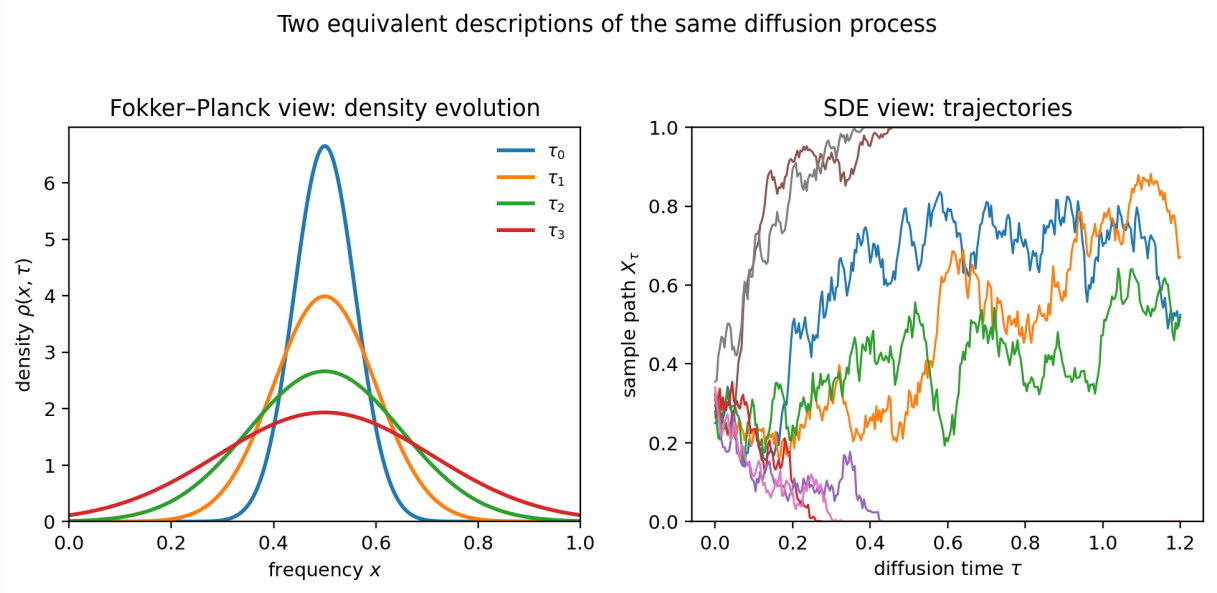

図 1　Fokker–Planck form と SDE form の違い．左は平均がほぼ同じまま広がっていく density $\rho(x,\tau)$ の模式図，右は同じ diffusion process の sample path の例である．Fokker–Planck equation は「分布全体がどう動くか」を，SDE は「個々の trajectory がどう動くか」を見るのに向いている．

である．したがって

$$\mathbb{E}[X_{t+1} \mid X_t = x] = x,$$

$$\mathrm{Var}(X_{t+1} \mid X_t = x) = \frac{x(1-x)}{N}$$

である．よって

$$\mathbb{E}[\Delta X_t \mid X_t = x] = 0,$$

$$\mathrm{Var}(\Delta X_t \mid X_t = x) = \frac{x(1-x)}{N}$$

となる．

テスト関数 f に対して Taylor 展開を用いると，

$$\mathbb{E}[f(X_{t+1}) - f(X_t) \mid X_t = x] = f'(x)\mathbb{E}[\Delta X_t \mid x] + \frac{1}{2}f''(x)\mathbb{E}[(\Delta X_t)^2 \mid x] + o\left(\frac{1}{N}\right).$$

中立では第 1 項が 0 なので，

$$\mathbb{E}[f(X_{t+1}) - f(X_t) \mid X_t = x] = \frac{1}{2N}x(1-x)f''(x) + o\left(\frac{1}{N}\right).$$

時間を $\tau = t/N$ とスケールすると，generator は

$$\mathcal{L}f(x) = \frac{1}{2}x(1-x)f''(x)$$

となる．したがって，中立な Wright–Fisher diffusion は

$$\mathrm{d}X_\tau = \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

である．同じ process を密度 $\rho(x,\tau)$ の時間発展として書くと，Fokker–Planck equation は

$$\frac{\partial \rho}{\partial \tau} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{x(1-x)\rho\}$$

となる．mutation がない場合は $x = 0, 1$ が absorbing boundary なので，Fokker–Planck form では境界条件の扱いも重要になる．

この SDE の読み方

右辺には deterministic term がなく，Brownian motion によるノイズ項だけがある．これは「平均的にはどちらにも進まないが，ランダムには揺らぐ」という中立モデルの性質を表している．ノイズの大きさは $\sqrt{x(1-x)}$ なので，頻度が中間付近のときに揺らぎが大きく，境界 $0, 1$ に近づくと小さくなる．

この diffusion process では，$x = 0$ と $x = 1$ が absorbing boundary である．mutation がない場合，最終的にはどちらかのタイプに固定する．中立な場合，初期頻度が x ならタイプ A の fixation probability は

$$\mathbb{P}(A \text{ fixes} \mid X_0 = x) = x$$

である．

3.3 selection がある場合の diffusion approximation

selection coefficient が固定値 s のままでは，1 世代あたりの平均変化が一般に $O(1)$ となる．diffusion approximation で genetic drift と selection を同じ時間スケールで見るためには，weak selection

$$s = \frac{\sigma}{N}$$

を仮定する．

weak selection という仮定の意味

ここでの「weak」は，selection が無視できるという意味ではない．1 世代あたりの selection effect を $O(1/N)$ にして，genetic drift による揺らぎと同じ長い時間スケールで比較できるようにするためのスケーリングである．この設定では，selection による小さな偏りが多くの世代にわたって積み重なる様子を，diffusion process の deterministic term として表せる．

このとき，親としてタイプ A が選ばれる確率は

$$q_s(x) = \frac{(1+\sigma/N)x}{1+\sigma x/N}$$

である．N が大きいとき，

$$q_s(x) = x + \frac{\sigma}{N}x(1-x) + O\left(\frac{1}{N^2}\right)$$

である．したがって

$$\mathbb{E}[\Delta X_t \mid X_t = x] = \frac{\sigma}{N} x(1-x) + O\left(\frac{1}{N^2}\right)$$

となる．variance の主要項は neutral の場合と同じで，

$$\mathrm{Var}(\Delta X_t \mid X_t = x) = \frac{x(1-x)}{N} + o\left(\frac{1}{N}\right)$$

である．よって，selection を含む Wright–Fisher diffusion は次である．

selection を含む Wright–Fisher diffusion

$$\mathrm{d}X_\tau = \sigma X_\tau (1 - X_\tau)\,\mathrm{d}\tau + \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

Fokker–Planck form では，密度 $\rho(x,\tau)$ は

$$\frac{\partial \rho}{\partial \tau} = -\frac{\partial}{\partial x}\{\sigma x(1-x)\rho\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{x(1-x)\rho\}$$

を満たす．SDE form の deterministic term $\sigma x(1-x)$ は，Fokker–Planck form では密度を押し流す transport term として現れる．

selection は deterministic term として現れる

weak selection $s = \sigma/N$ のもとでは，1 世代あたりの小さな偏りが，rescale された時間 $\tau = t/N$ で

$$\sigma x(1-x)$$

という deterministic term として残る．この deterministic term を使うと，有利な変異が最終的に fix する確率も解析できる．fixation probability の導出は次の小節でまとめる．

3.4 fixation probability: diffusion approximation による導出

ここでは，mutation がない場合を考える．一度 $X = 0$ または $X = 1$ に到達すると，その状態から抜け出せないので，0 と 1 は absorbing boundary である．初期頻度 $X_0 = x$ から出発したタイプ A の fixation probability を

$$\pi(x) = \mathbb{P}(A \text{ fixes} \mid X_0 = x)$$

と書く．

fixation probability は，backward equation

$$\mathcal{L}\pi(x) = 0, \qquad \pi(0) = 0, \qquad \pi(1) = 1$$

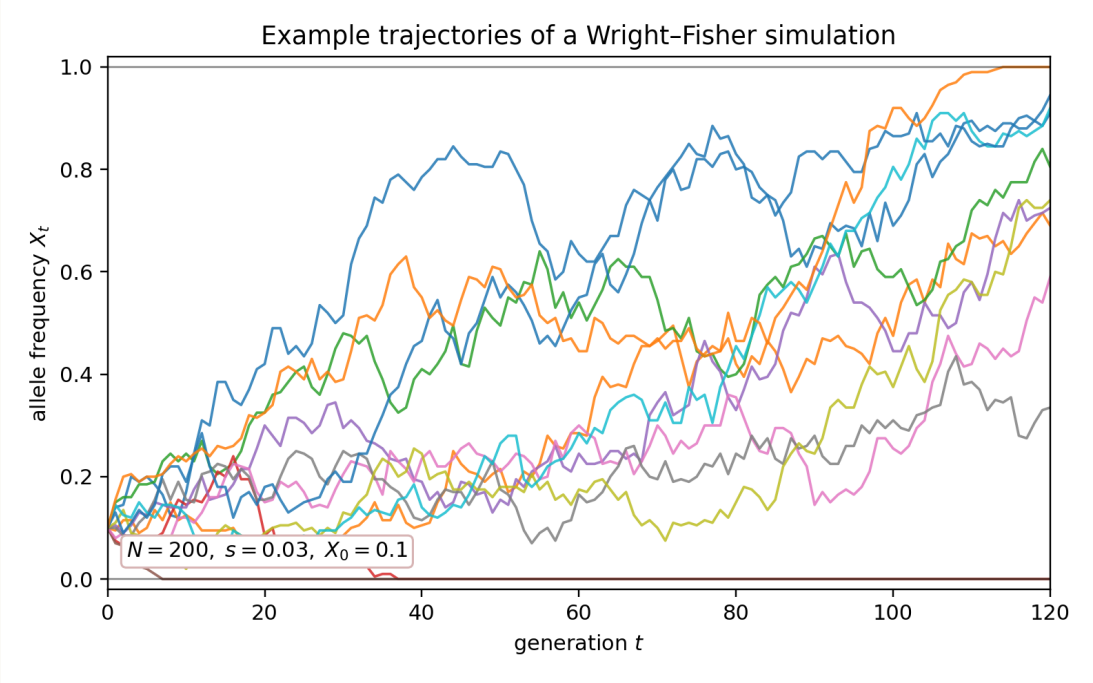

図 2　Type based な Wright–Fisher simulation の例．$N = 200$, $s = 0.03$, $X_0 = 0.1$ として独立な trajectory を複数本シミュレートした．平均的には有利変異 A が増えやすいが，各 run では genetic drift のために大きならばらつきがある．初期に消失する run もあり，一方で高頻度まで増える run もある．

を満たす．weak selection の diffusion approximation では

$$\mathcal{L}f(x) = \sigma x(1-x)f'(x) + \frac{1}{2}x(1-x)f''(x)$$

であるから，$0 < x < 1$ では

$$\sigma x(1-x)\pi'(x) + \frac{1}{2}x(1-x)\pi''(x) = 0$$

となる．$x(1-x) > 0$ で割ると

$$\pi''(x) + 2\sigma\pi'(x) = 0$$

である．ここで $g(x) = \pi'(x)$ とおくと

$$g'(x) + 2\sigma g(x) = 0$$

なので

$$g(x) = Ce^{-2\sigma x}$$

である．積分して

$$\pi(x) = A + Be^{-2\sigma x}$$

と書ける．境界条件 $\pi(0) = 0$，$\pi(1) = 1$ を使うと

$$\pi(x) = \frac{1 - e^{-2\sigma x}}{1 - e^{-2\sigma}}$$

を得る．neutral limit $\sigma \to 0$ では

$$\pi(x) = x$$

に戻る．

実際の世代ごとの selection coefficient を s と書くと，この式は $\sigma = Ns$ と見ればよい．つまり，近似的に

$$\pi(x) = \frac{1 - e^{-2Nsx}}{1 - e^{-2Ns}}$$

である．特に，1 個体だけの新しい変異から始まる場合は $x = 1/N$ なので

$$\pi_1 = \pi\left(\frac{1}{N}\right) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \simeq \frac{2s}{1 - e^{-2Ns}}$$

となる．さらに $s \ll 1$ かつ $Ns \gg 1$ なら

$$\pi_1 \simeq 2s$$

である．この $\pi_1 \simeq 2s$ という結果は、
bfseries Haldaneś formula としても知られる。

mutation なし，haploid，タイプ A の relative fitness が $1 + s$ のとき，diffusion approximation から次を得る．

fixation probability の基本公式

$$\pi(x) \simeq \frac{1 - e^{-2Nsx}}{1 - e^{-2Ns}}$$

$$x = \frac{1}{N}, \quad s \ll 1, \quad Ns \gg 1 \quad \Longrightarrow \quad \pi_1 \simeq 2s.$$

中立なら fixation probability は初期頻度そのもの，つまり 1 個体なら $1/N$ である．

3.5 fixation probability: branching process による導出

有利な変異が 1 個体だけ生じた直後を考える．この段階では変異型は非常にまれなので，変異型同士の競争や集団全体の頻度変化はまだ無視できる．そのため，変異型の初期の fate を branching process で近似できる．ここで知りたいのは，変異型が初期の stochastic extinction を逃れて，十分な個体数まで増える確率である．大きな集団で $Ns \gg 1$ のとき，この establishment probability は fixation probability とほぼ同じオーダーになる．

ここでは，連続時間の birth–death process，あるいは Moran model の初期段階のように考える．変異型が 1 個体だけいるとき，次にその系統に関係する有効な event は，

- 変異型が出生して，変異型が 1 個体から 2 個体になる，
- 変異型が死亡して，変異型が 1 個体から 0 個体になる，

のどちらかであると近似する．有利変異では，出生側が少し起こりやすい．Wright–Fisher diffusion の fixation probability で用いた selection coefficient s とそろえるため，初期段階の有効な birth rate と death rate を

$$b = 1 + s, \qquad d = 1 - s$$

と置く．すると，最初の event が出生である確率と死亡である確率は

$$p_+ = \frac{b}{b+d} = \frac{1+s}{2}, \qquad p_- = \frac{d}{b+d} = \frac{1-s}{2}$$

である．ここで s は十分小さい正の数とする．

1 個体から始まった変異型系統の establishment probability を π と書く．最初の event で死亡すれば，系統はそこで絶滅するので寄与は 0 である．最初の event で出生すれば，変異型は 2 個体になる．branching process 近似では，この 2 個体の子孫系統は互いに独立に振る舞うと考える．したがって，2 個体系統のうち少なくとも一つが establish する確率は

$$1 - (1-\pi)^2 = 2\pi - \pi^2$$

である．以上より，π は recursion relation

$$\pi = p_+\{1 - (1-\pi)^2\} + p_- \cdot 0 = \frac{1+s}{2}(2\pi - \pi^2)$$

を満たす．$\pi > 0$ の解を求めるため，両辺を π で割ると

$$1 = (1+s)\left(1 - \frac{\pi}{2}\right)$$

である．したがって

$$\pi = \frac{2s}{1+s}$$

を得る．$s \ll 1$ なら

$$\pi \simeq 2s$$

である．

> **branching process による $2s$ の導出**
>
> $$\pi = \frac{1+s}{2}\{1 - (1-\pi)^2\} \quad \Longrightarrow \quad \pi = \frac{2s}{1+s} \simeq 2s.$$

この導出で使っているのは，fixation そのものを最後まで追うというより，「まれな有利変異が初期の genetic drift による消失を逃れる確率」を計算するという考え方である．変異型が十分に増えると，その後は deterministic な selection の効果で高頻度まで増えやすい．そのため，大きな集団で $s \ll 1$，$Ns \gg 1$ のとき，この establishment probability と 1 個体からの fixation probability はどちらも $2s$ で近似される．

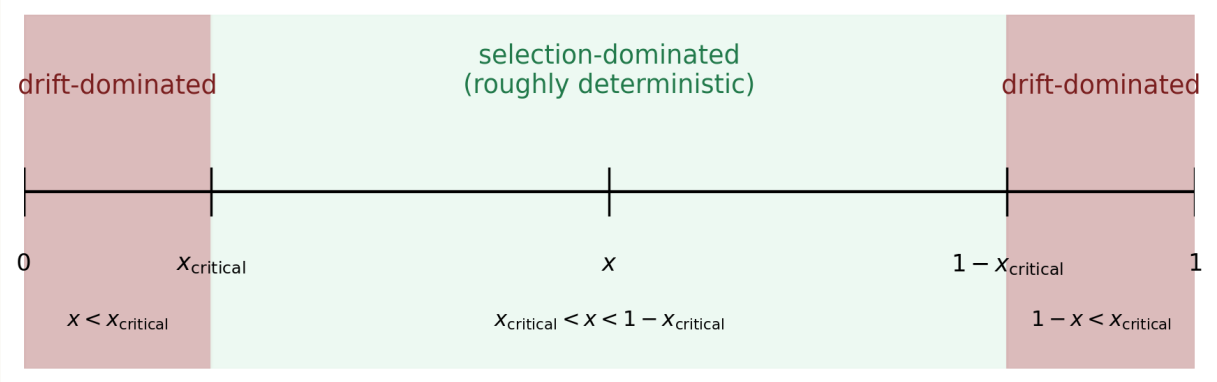

図 3　Drift–selection boundary の模式図．低頻度側 $x < x_{\mathrm{critical}}$ と高頻度側 $1 - x < x_{\mathrm{critical}}$ では genetic drift が支配的であり，中間の領域 $x_{\mathrm{critical}} < x < 1 - x_{\mathrm{critical}}$ では selection が支配的になる．$P_{\mathrm{fix}} \simeq 2s$ は，1 個体の有利変異が左側の drift-dominated 領域を抜ける確率として読むことができる．

Drift–selection boundary

$P_{\mathrm{fix}} \simeq 2s$ という近似は，「有利変異がまず drift-dominated な低頻度領域を抜ける確率」として読むと分かりやすい．変異型が m 個体あるとき，頻度は $x = m/N$ である．初期段階では，selection による bias が genetic drift によるランダムな消失と同程度になる境界を

$$2sm \sim 1$$

で見積もることができる．したがって，critical copy number と critical frequency は

$$m_{\mathrm{critical}} \sim \frac{1}{2s}, \qquad x_{\mathrm{critical}} = \frac{m_{\mathrm{critical}}}{N} \sim \frac{1}{2Ns}$$

である．頻度が

$$x < x_{\mathrm{critical}}$$

の領域では，変異型はまだ少なすぎるため，その fate は主に genetic drift に支配される．同様に，固定に非常に近い側では

$$1 - x < x_{\mathrm{critical}}$$

なら，野生型側の copy number が小さいため，境界付近の挙動は drift-dominated になる．一方で

$$x_{\mathrm{critical}} < x < 1 - x_{\mathrm{critical}}$$

の中間領域では，selection の効果が genetic drift より大きく，軌道はおおよそ deterministic に

$$\frac{\mathrm{d}x}{\mathrm{d}t} \simeq sx(1 - x)$$

に従うと考えられる．したがって，大きな集団で $Ns \gg 1$ のとき，$P_{\mathrm{fix}} \simeq 2s$ は，「1 個体の有利変異が低頻度側の drift-dominated 領域を抜けて establish する確率」を表している，と解釈できる．

3.6 mutation がある場合の diffusion approximation

mutation も genetic drift と同じ時間スケールで見るため，

$$u = \frac{\mu}{N}, \qquad v = \frac{\nu}{N}$$

とおく．

mutation rate もスケールする理由

mutation rate を固定値のままにすると，1 世代あたりの頻度変化が大きくなり，diffusion approximation で見ている t/N の時間スケールでは mutation の効果が強すぎることがある．$u = \mu/N$，$v = \nu/N$ と置くことで，mutation による流入・流出も，genetic drift や weak selection と同じスケールで比較できる．

neutral で mutation だけがある場合，次世代個体がタイプ A である確率は

$$p(x) = (1-u)x + v(1-x)$$

である．したがって

$$p(x) - x = -ux + v(1-x) = \frac{1}{N}\{\nu(1-x) - \mu x\}$$

である．よって mutation による deterministic term は

$$b_{\mathrm{mut}}(x) = \nu(1-x) - \mu x$$

となる．

mutation を含む neutral な Wright–Fisher diffusion は次である．

> **mutation を含む neutral Wright–Fisher diffusion**
>
> $$\mathrm{d}X_\tau = \{\nu(1-X_\tau) - \mu X_\tau\}\,\mathrm{d}\tau + \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

対応する Fokker–Planck equation は

$$\frac{\partial \rho}{\partial \tau} = -\frac{\partial}{\partial x}[\{\nu(1-x) - \mu x\}\rho] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{x(1-x)\rho\}$$

である．この形で見ると，mutation は density を境界から内部へ押し戻す deterministic な方向づけとして働くことが分かる．

双方向 mutation $\mu > 0, \nu > 0$ がある場合，0 と 1 は absorbing boundary ではなくなる．neutral で双方向 mutation があるときの stationary density は

$$\rho_*(x) = \frac{1}{B(2\nu, 2\mu)} x^{2\nu-1}(1-x)^{2\mu-1}$$

である．ここで $B(2\nu, 2\mu)$ は density を積分して 1 にするための normalization constant である．これは Beta distribution $\mathrm{Beta}(2\nu, 2\mu)$ であり，$[0,1]$ 上の連続分布としてアリル頻度の stationary density を表している．

3.7 selection と mutation がある場合

weak selection と weak mutation

$$s = \frac{\sigma}{N}, \qquad u = \frac{\mu}{N}, \qquad v = \frac{\nu}{N}$$

を仮定する．selection 後の親の A 頻度は

$$q_s(x) = \frac{(1+\sigma/N)x}{1+\sigma x/N}$$

であり，mutation 後に次世代個体が A である確率は

$$p_{s,u,v}(x) = (1-u)q_s(x) + v\{1-q_s(x)\}.$$

展開すると，

$$p_{s,u,v}(x) = x + \frac{1}{N}\left\{\sigma x(1-x) + \nu(1-x) - \mu x\right\} + O\left(\frac{1}{N^2}\right).$$

したがって，diffusion approximation は次である．

> **selection と mutation を含む Wright–Fisher diffusion**
>
> $$\mathrm{d}X_\tau = \left\{\sigma X_\tau(1-X_\tau) + \nu(1-X_\tau) - \mu X_\tau\right\}\mathrm{d}\tau + \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

同じ process を Fokker–Planck form で書くと，密度 $\rho(x,\tau)$ は

$$\frac{\partial \rho}{\partial \tau} = -\frac{\partial}{\partial x}\left[\left\{\sigma x(1-x) + \nu(1-x) - \mu x\right\}\rho\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)\rho\right].$$

この式は，selection と mutation が density をどちらへ運ぶか，genetic drift が density をどのように広げるかを表している．

4 実装方法

Wright–Fisher model の実装には，大きく分けて **individual based** な実装と **type based** な実装がある．どちらを使うべきかは，個体ごとの情報を保持したいか，タイプ頻度だけを追跡すればよいかによって決まる．

実装方法を分けて考える意義

理論式では I_t や X_t だけを扱うことが多いが，シミュレーションでは「何を状態として保存するか」を決める必要がある．頻度だけに興味があるなら type based 実装で十分であり，高速である．一方，個体ごとの系譜，空間位置，表現型，多遺伝子座などを扱いたいなら，individual based 実装が必要になる．

4.1 Individual based 実装

Individual based 実装では，長さ N の配列で各個体のタイプを持つ．たとえばタイプ A を 1，タイプ a を 0 として，

$$\boldsymbol{z}_t = (z_{1,t}, \ldots, z_{N,t}), \qquad z_{k,t} \in \{0, 1\}$$

を保持する．

Wright–Fisher 更新は以下のように行う．

1. 各個体に fitness を割り当てる．
2. fitness に比例して親を N 回復元抽出する．
3. 親のタイプを子にコピーする．
4. mutation を適用する．

Python 風の擬似コードは次の通りである．

```
import numpy as np

def wf_individual_based(types, s=0.0, u=0.0, v=0.0, rng=None):
    if rng is None:
        rng = np.random.default_rng()

    N = len(types)

    weights = np.where(types == 1, 1.0 + s, 1.0)
    probs = weights / weights.sum()

    parent_index = rng.choice(N, size=N, replace=True, p=probs)
    offspring = types[parent_index].copy()

    mut_A_to_a = (offspring == 1) & (rng.random(N) < u)
    mut_a_to_A = (offspring == 0) & (rng.random(N) < v)

    offspring[mut_A_to_a] = 0
    offspring[mut_a_to_A] = 1

    return offspring
```

Individual based 実装の利点は，個体ごとの情報を保持できることである．たとえば，系譜，空間位置，複数遺伝子座，個体ごとに異なる増殖率などを自然に扱える．一方で，集団サイズ N が大きいと計算コストが高くなる．

Individual based で見えるもの

Individual based 実装は，単に頻度を追うだけなら過剰に見えることがある．しかし，「どの個体がどの個体の子孫か」，「どの場所にいる個体が増えたか」，「複数の形質が同じ個体内でどう組み合わさるか」を見たい場合には有用である．つまり，個体の履歴や内部状態が研究上の問いに含まれるなら，この実装が自然である．

4.2 Type based 実装

Type based 実装では，各タイプの個体数だけを追跡する．bi-allelic model なら，状態は I_t だけで十分である．

現在 $I_t = i$，$x = i/N$ とする．selection と mutation を含む Wright–Fisher model では，

$$q_s(x) = \frac{(1+s)x}{1+sx}$$

とし，

$$p_{s,u,v}(x) = (1-u)q_s(x) + v\{1-q_s(x)\}$$

を計算する．その後，

$$I_{t+1} \sim \mathrm{Bin}(N, p_{s,u,v}(x))$$

をサンプリングすればよい．

Python 風の擬似コードは次の通りである．

```
import numpy as np

def wf_type_based(i, N, s=0.0, u=0.0, v=0.0, rng=None):
    if rng is None:
        rng = np.random.default_rng()

    x = i / N
    q = (1.0 + s) * x / (1.0 + s * x)
    p = (1.0 - u) * q + v * (1.0 - q)
    p = min(max(p, 0.0), 1.0)

    return rng.binomial(N, p)
```

Type based 実装は，bi-allelic model や少数タイプのモデルでは非常に効率がよい．特に多数の反復シミュレーションを行う場合，individual based 実装よりもかなり高速である．

Type based で十分な問い

fixation probability，mean fixation time，頻度分布，selection coefficient を変えたときの軌道の違いなど，タイプの個体数だけで答えられる問いでは，type based 実装が第一選択になる．この場合，集団内の「誰が誰の子か」は捨てて，各世代のタイプ数だけを効率よくサンプリングする．

4.3 bi-allelic SDE の実装: selection ありの Itô 離散化

ここまでの individual based 実装と type based 実装は，有限集団の Wright–Fisher model を直接シミュレートする方法である．一方，集団サイズが十分大きいとみなせる場合には，diffusion approximation として得られた SDE を直接シミュレートすることもできる．ここでは，mutation なし・selection ありの bi-allelic Wright–Fisher diffusion

$$\mathrm{d}X_\tau = \sigma X_\tau(1-X_\tau)\,\mathrm{d}\tau + \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

を実装する．

この SDE は Itô SDE なので，最も基本的には Euler–Maruyama method で離散化する．時間刻みを $\Delta\tau$ とし，$\xi_n \sim \mathcal{N}(0,1)$ を独立にサンプリングすると，次のように書ける．ここで $\mathcal{N}(m,\eta^2)$ は mean m，variance η^2 の normal distribution であり，$\mathcal{N}(0,1)$ は standard normal distribution である．

> **selection あり bi-allelic SDE の Itô 離散化**
>
> $$X_{n+1} = X_n + \sigma X_n(1-X_n)\Delta\tau + \sqrt{X_n(1-X_n)\Delta\tau}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0,1).$$

これは Brownian increment を

$$\Delta W_n = W_{\tau_{n+1}} - W_{\tau_n} \sim \mathcal{N}(0,\Delta\tau)$$

と表し，

$$\Delta W_n = \sqrt{\Delta\tau}\,\xi_n$$

と書いたものである．

元の Wright–Fisher model で selection coefficient を s と書いていた場合，diffusion approximation の weak selection では

$$s = \frac{\sigma}{N}, \qquad \text{つまり} \qquad \sigma = Ns$$

である．また，diffusion time は $\tau = t/N$ なので，t 世代分と比較したいなら $\tau = t/N$ まで SDE を進める．

Python 風の擬似コードは次の通りである．

```
import numpy as np
```

```python
def wf_sde_selection(x0, sigma, dtau, n_steps, rng=None):
    """
    Simulate the selection-only bi-allelic Wright-Fisher diffusion

        dX = sigma X(1-X) dtau + sqrt(X(1-X)) dW

    by Euler-Maruyama discretization.
    Here dtau is the time step in diffusion time tau = t / N.
    """
    if rng is None:
        rng = np.random.default_rng()

    xs = np.empty(n_steps + 1)
    xs[0] = x0

    for n in range(n_steps):
        x = xs[n]

        # Without mutation, x=0 and x=1 are absorbing boundaries.
        if x <= 0.0:
            xs[n + 1:] = 0.0
            break
        if x >= 1.0:
            xs[n + 1:] = 1.0
            break

        deterministic = sigma * x * (1.0 - x)
        diffusion = np.sqrt(x * (1.0 - x))
        xi = rng.normal(0.0, 1.0)

        x_next = x + deterministic * dtau + diffusion * np.sqrt(dtau) * xi

        # Euler-Maruyama may slightly overshoot [0, 1].
        # For small dtau this is rare; here we project back to the interval.
        x_next = min(max(x_next, 0.0), 1.0)
        xs[n + 1] = x_next

    return xs
```

SDE 実装で注意すること

この実装は，有限集団の binomial sampling をそのまま再現するものではなく，$N \to \infty$ の diffusion approximation を数値的に解くものである．そのため，N が小さい場合や，fixation 直前の境界付近を高精度に扱いたい場合には，元の type based Wright–Fisher simulation の方が自然なこともある．また，Euler–Maruyama method では有限の刻み幅のために $[0, 1]$ の外へ少し出ることがあるの

で，上のコードでは簡単のため projection している．より丁寧に扱う場合は，刻み幅を小さくする，boundary treatment を工夫する，あるいは元の discrete model で確認するのがよい．

4.4 多タイプへの拡張

タイプが K 個ある場合，各タイプの個体数を

$$\boldsymbol{n}_t = (n_{1,t}, \dots, n_{K,t}), \qquad \sum_{k=1}^{K} n_{k,t} = N$$

と書く．タイプ k の fitness を w_k とする．selection 後に親がタイプ k である確率は

$$q_k = \frac{w_k n_k}{\sum_{\ell=1}^{K} w_\ell n_\ell}$$

である．mutation matrix を $M = (M_{k\ell})$ とし，

$$M_{k\ell} = \mathbb{P}(\text{親タイプ } k \text{ から子タイプ } \ell)$$

と定義する．すると，次世代個体がタイプ ℓ である確率は

$$p_\ell = \sum_{k=1}^{K} q_k M_{k\ell}$$

である．したがって，次世代の個体数ベクトルは

$$\boldsymbol{n}_{t+1} \sim \mathrm{Mult}(N; p_1, \dots, p_K)$$

に従う．ここで $\mathrm{Mult}(N; p_1, \dots, p_K)$ は multinomial distribution である．これは，K 個のタイプのうち 1 回の試行でタイプ k が選ばれる確率を p_k とし，それを N 回独立に繰り返したとき，各タイプが何個体ずつ現れるかを表す分布である．bi-allelic model で使った binomial distribution を，多タイプに拡張したものと考えればよい．

4.5 実装上の注意

- **時間スケールを明示する**：Wright–Fisher では 1 step が 1 世代である．Moran では 1 step が 1 birth–death event である．
- **weak selection・weak mutation のスケーリングに注意する**：diffusion approximation と対応させるなら，典型的には $s = \sigma/N$，$u = \mu/N$，$v = \nu/N$ とする．
- **boundary を確認する**：mutation がない場合，$i = 0$ と $i = N$ は absorbing state である．双方向 mutation がある場合は absorbing state ではない．
- **replicate 数を十分に取る**：単一のシミュレーション軌道は大きく揺らぐため，fixation probability や mean fixation time を推定するには多数の replicate が必要である．

- **type based と individual based を使い分ける**：頻度だけで十分なら type based，個体ごとの履歴や形質が必要なら individual based が適している．

5 まとめ

Wright–Fisher model は，有限集団におけるアリル頻度変化を記述する基本的な確率モデルである．Moran model とは更新単位が異なるが，neutral な場合には時間を rescale すると同じ Wright–Fisher diffusion に収束する．したがって，Moran model と Wright–Fisher model を比べるときは，step 数そのものではなく時間スケールをそろえることが重要である．

neutral な Wright–Fisher model では，頻度の期待値は変わらないが，有限集団サイズに由来する variance

$$\mathrm{Var}(X_{t+1} \mid X_t = x) = \frac{x(1-x)}{N}$$

が生じる．これが genetic drift である．selection は平均的な頻度変化を生み，mutation はタイプ間の流入・流出を生む．

mutation がない場合，fixation probability は backward equation から導出できる．有利変異が 1 個体から始まる場合，$s \ll 1$ かつ $Ns \gg 1$ なら

$$\pi_1 \simeq 2s$$

である．同じ $2s$ は，まれな有利変異の初期増殖を branching process で近似しても得られる．

大集団極限では，weak selection・weak mutation のもとで，Wright–Fisher model は

$$\mathrm{d}X_\tau = \{\sigma X_\tau(1-X_\tau) + \nu(1-X_\tau) - \mu X_\tau\}\,\mathrm{d}\tau + \sqrt{X_\tau(1-X_\tau)}\,\mathrm{d}W_\tau$$

という diffusion process で近似できる．この SDE と等価に，対応する Fokker–Planck equation を使えば，頻度分布そのものの時間発展を記述できる．

実装では，bi-allelic model なら個体数 I_t のみを追跡する type based 実装が効率的である．大きな集団の diffusion approximation を直接見たい場合には，selection ありの bi-allelic SDE を Itô 離散化してシミュレートできる．一方，個体ごとの履歴，空間構造，複数形質などを扱う場合には individual based 実装が有用である．

参考文献

- Ewens, W. J. *Mathematical Population Genetics*. Springer.
- Durrett, R. *Probability Models for DNA Sequence Evolution*. Springer.
- Etheridge, A. *Some Mathematical Models from Population Genetics*. Springer.
- Wakeley, J. *Coalescent Theory: An Introduction*. Roberts & Company.