Populációgenetikai modellek

Szakdolgozat
Matematika BSc

Készítette: Gerencsér Máté
Témavezető: Csiszár Villő
adjunktus

Eötvös Loránd Tudományegyetem
Természettudományi Kar
Budapest, 2010

Tartalomjegyzék

1. Előszó

A populációgenetika a populációk genetikai összetételével, annak változásával, illetve a genetikai szerkezetet meghatározó folyamatokkal foglalkozik. Általában egy adott lókuszban vizsgáljuk a gének előforduló típúsait, az allélokat. Ha az egyes allélokat egy valószínűségi változó értékkészletének tekintjük, akkor az egymást követő generációk egy sztochasztikus folyamatot alkotnak. A matematikai modellezés során a generációk közötti kapcsolatra teszünk olyan feltételezéseket, amikről azt gondoljuk, hogy nem állnak távol a valóságtól, ugyanakkor matematikailag kezelhetővé válik a folyamat.

A dolgozat során végig haploid populációkkal foglalkozom, azaz olyanokkal, ahol minden egyed egy alléllal rendelkezik (ezért azonosítjuk is vele), továbbá nemek nincsenek megkülönböztetve. A genetikai változásokat általában négy tényező befolyásolja: a mutáció, a véletlen okozta genetikai sodródás, a természetes szelekció, és a populációk közti génáramlás. A tárgyalt modellekben az első kettőt fogjuk figyelembe venni.

Dolgozatom második fejezetében [5] alapján a Markov-folyamatokról későbbiekhez szükséges tudnivalókat foglalom össze. A harmadik és negyedik fejezetben két klasszikus modellt tárgyalok. E kettőt is magában foglalja az ötödik fejezetben bemutatott Cannings-modell, aminek egy szép jellemzését fogjuk látni. Végül a hatodik fejezetben egy, a korábbiaktól jelentősen eltérő módszert mutatok be a populációk vizsgálatára. A dolgozatban főleg [3]-ra támaszkodtam, ezen kívül az ötödik fejezetben [1]-t, a hatodikban [7]-t használtam fel jelentősen.

Köszönetet szeretnék mondani témavezetőmnek, Csiszár Villőnek, aki értékes tanácsaival nagyban segítette a szakdolgozat elkészültét.

2. Markov-folyamatok

2.1. Alapfogalmak, Markov-láncok

*Sztochasztikus folyamat*on valószínűségi változók egy $\{X_t, t \in T\}$ családját értjük. A T halmaz a folyamat *paramétertartománya*, erre sokszor mint időre tekintünk. Azt a – természetesen nem egyértelműen meghatározott, de általában természetesen adódó – S halmazt pedig, amiben az X_t változók az értékeiket felveszik, *állapottér*nek nevezzük. Egy sztochasztikus folyamat *Markov-folyamat*, ha amennyiben a folyamat jelenlegi állapotát ismerjük, a jövőbeli állapot nem függ a korábbi állapotoktól, azaz $t_1 < t_2 < \ldots < t_n < t$ esetén $P(X_t \in A | X_{t_1} = x_1, X_{t_2} = x_2, ..., X_{t_n} = x_n) = P(X_t \in A | X_{t_n} = x_n)$. *Markov-lánc*nak olyan Markov-folyamatot hívunk, aminek állapottere megszámlálható, paramétertartománya pedig $T = \{0,1,2,...\}$. Most csak olyan Markov-láncokkal foglalkozunk, ahol $S = \{0,1,2,...,N\}$ $(N \leq \infty)$; az $X_t = i$ eseményre úgy is hivatkozunk, hogy X_t az i állapotban van.

Az *egy lépéses átmenetvalószínűségeket* a következőképpen definiáljuk: $P_{ij}^{\{n\}} = P(X_{n+1} = j | X_n = i)$. *Stacionárius* átmenetvalószínűségekről beszélünk, ha ez nem függ n-től. Ezt a továbbiakban föltesszük, ekkor elhagyjuk a felső indexet. A $P = \{P_{ij}\}$ $(N+1) \times (N+1)$ méretű mátrixot *átmenetmátrix*nak nevezzük. Ha a P_{ij} valószínűségeket és X_0 kezdeti értékét (vagy eloszlását) megadjuk, azzal az egész folyamatot meghatároztuk, azaz ezek alapján minden $P(X_{t_1} = x_1, X_{t_2} = x_2, ..., X_{t_n} = x_n)$ valószínűség kiszámolható. Az n lépéses átmenetvalószínűség $P_{ij}^n = P(X_{m+n} = j | X_m = i)$; az n lépéses átmenetmátrix $P^{(n)} = \{P_{ij}^n\}$. Fennáll a következő: $P_{ij}^n = \sum_{k=0}^{N} P_{ik}^r P_{kj}^s$, ha $r + s = n$, ebből pedig már adódik $P^{(n)} = P^n$.

A j állapot az i állapotból *elérhető*, ha van olyan n, hogy $P_{ij}^n > 0$. Az i és j állapotok érintkeznek, jelölésben $i \leftrightarrow j$, ha j elérhető i-ből és i is j-ből. Ez ekvivalenciareláció, az ekvivalenciaosztályok az állapottér *osztályai*. A Markov-lánc *irreducibilis*, ha egyetlen osztály van. Egy i állapot *rekurrens*, ha $\sum_{n=1}^{\infty} P_{ii}^n = \infty$. Ez ekvivalens azzal, hogy az i állapotból indulva a folyamat 1 valószínűséggel véges sok lépésben visszatér az i állapotba. Ha egy állapot nem rekurrens, akkor *tranziens*nek nevezzük. Ha egy állapot rekurrens, akkor minden vele egy osztályban levő állapot is az. Egy i állapot $d(i)$ *periódus*a azon n számok legnagyobb közös osztója, amikre $P_{ii}^n > 0$. Ha

nem létezik ilyen n, akkor $d(i) = 0$. Ha $i \leftrightarrow j$, akkor $d(i) = d(j)$. A Markov-lánc, vagy annak egy osztálya *aperiodikus*, ha minden állapotának periódusa 1. Ha egy aperiodikus, rekurrens osztálybeli i-re $\lim_{n\to\infty} P_{ii}^n =: \phi_i > 0$, akkor $\phi_j > 0$ minden vele egy osztályban levő j állapotra. Ekkor az osztályt *pozitív rekurrens*nek, ha pedig $\phi_i = 0$ minden osztálybeli i-re, akkor *nulla rekurrens*nek nevezzük.

2.1. Állítás. *Legyenek egy pozitív rekurrens, aperiodikus osztály állapotai* 0,1, ..., M. *Ekkor*

$$\lim_{n\to\infty} P_{jj}^n = \phi_j = \sum_{i=0}^{M} \phi_i P_{ij}, \quad \sum_{j=0}^{M} \phi_j = 1,$$

és a ϕ_j *számok egyértelmű nemnegatív megoldásai a*

$$\phi_j = \sum_{i=0}^{M} \phi_i P_{ij}, \quad \sum_{j=0}^{M} \phi_j = 1$$

egyenletrendszernek.

Ha a fenti egyenletrendszert az egész Markov-láncra írjuk fel (azaz lényegében M helyére N-et írunk), akkor a nemnegatív megoldásokat *stacionárius eloszlás*nak nevezzük. Ez tehát egyfajta egyensúlyi helyzet: ha X_0 eloszlása egy ϕ stacionárius eloszlás, akkor minden X_t, $t > 0$ eloszlása is ϕ lesz. Irreducibilis, véges állapotterű Markov-láncnak mindig egyértelműen létezik staconárius eloszlása.

Egy Markov-lánc *megfordítható*, ha van olyan $\pi = (\pi_0, \pi_1, ...\pi_N)$, hogy

$$\pi_i P_{ij} = \pi_j P_{ji}.$$

Ilyen tulajdonságú π-t 1-re normálva stacionárius eloszlást kapunk.

Jelölje Tr a tranziens állapotok halmazát, C pedig legyen egy rekurrens osztály. Jelölje továbbá $\pi_i(C)$ annak a valószínűségét, hogy i-ből indulva a folyamat belép a C osztályba. Miután belépett, kilépni már nem tud, így ezt C-ben való *elnyelődés*nek is nevezzük. A $\pi_i(C)$ számokra teljesül

$$\pi_i(C) = \sum_{j=0}^{N} P_{ij}\pi_j(C), \tag{1}$$

és ha a $w_i = \sum_{j \in Tr} P_{ij} w_j$ egyenletrendszer egyetlen korlátos megoldása $w \equiv 0$, akkor $\pi_i(C)$ meghatározható mint a fenti egyenletrendszer egyetlen megoldása.

2.2. Kontinuáns mátrixok

2.2. Definíció. *Egy $A = \{a_{ij}\}$ mátrixot kontinuánsnak, másnéven tridiagonálisnak nevezünk, ha $|i-j| > 1$ esetén $a_{ij} = 0$.*

Kontinuáns átmenetmátrixok esetén használatosak a következő jelölések:

$$P_{ij} = \begin{cases} \mu_i, & j = i-1, \\ \lambda_i, & j = i+1, \\ 1 - \mu_i - \lambda_i, & j = i, \\ 0, & |i-j| > 1; \end{cases} \qquad (i,j = 0,1,2,...,n)$$

$$\rho_0 = 1, \quad \rho_i = \frac{\mu_1 \mu_2 \dots \mu_i}{\lambda_1 \lambda_2 \dots \lambda_i} \qquad (i = 1,2,...,n).$$

2.3. Állítás. *Tegyük fel, hogy a P kontinuáns átmenetmátrixra $\lambda_0 = \mu_n = 0$, a megfelelő Markov-lánc indulóállapota pedig legyen k. Ekkor*

$$P(\textit{elnyelődés 0-ban}) = 1 - P(\textit{elnyelődés n-ben}) = \frac{\sum_{i=k}^{n-1} \rho_i}{\sum_{i=0}^{n-1} \rho_i}.$$

Bizonyítás. Legyen a keresett valószínűség α_k, ekkor felírható:

$$\alpha_k = \lambda_k \alpha_{k+1} + \mu_k \alpha_{k-1} + (1 - \lambda_k - \mu_k)\alpha_k,$$

átrendezéssel

$$\alpha_k - \alpha_{k+1} = \frac{\mu_k}{\lambda_k}(\alpha_{k-1} - \alpha_k) = ... = \prod_{i=1}^{k} \frac{\mu_i}{\lambda_i}(\alpha_0 - \alpha_1) = \rho_k(1 - \alpha_1).$$

Ezt használva,

$$\alpha_k = \sum_{i=k}^{n-1} (\alpha_i - \alpha_{i+1}) = (1 - \alpha_1) \sum_{i=k}^{n-1} \rho_i.$$

$k = 1$ helyettesítéssel $(1 - \alpha_1) = \frac{1}{\sum_{i=0}^{n-1} \rho_i}$ adódik, így a fenti képlet pont az állítást adja. ■

2.4. Állítás. *Ha $\lambda_0 > 0$, $\mu_n > 0$ akkor a P átmenetmátrixhoz tartozó Markov-lánc irreducibilis, így létezik ϕ stacionárius eloszlás, amire*

$$\phi_k := P(X_t = k) = c\frac{\lambda_0\lambda_1 \dots \lambda_{k-1}}{\mu_1\mu_2 \dots \mu_k} \quad (k = 0,1, ..., n)$$

valamilyen c konstansra.

Bizonyítás. A stacionárius eloszlásra $\phi = \phi P$, speciálisan

$$\phi_0 = (1 - \lambda_0)\phi_0 + \mu_1\phi_1,$$

ahonnan $\phi_1 = \frac{\lambda_0}{\mu_1}\phi_0$, így az állítás $c = \phi_0$ választással igaz $k = 0,1$ esetén. Ha pedig már $(k-1)$-re és k-ra tudjuk, akkor a

$$\phi_k = \phi_{k-1}\lambda_{k-1} + \phi_{k+1}\mu_{k+1} + (1 - \lambda_k - \mu_k)\phi_k$$

egyenletbe behelyettesítve $(k+1)$-re is adódik az állítás, így indukcióval minden k-ra. A $c = \phi_0$ konstans a $\sum_{i=0}^{n} \phi_i = 1$ feltételből számolható ki. ■

Visszatérve a $\lambda_0 = 0$, $\mu_n = 0$ esetre, az előzőkhöz hasonló gondolatmenettel kaphatjuk meg az i állapotból indulva az elnyelődésig eltelt idő t_i várható értékét. Ha $t_{i,j}$ az i állapotból indulva a j állapotban töltött idő várható értéke, akkor nyilván $t_i = \sum_{j=1}^{n-1} t_{i,j}$. Tudjuk továbbá a következőt:

$$t_{i,j} = \mu_i t_{i-1,j} + \lambda_i t_{i+1,j} + (1 - \mu_i - \lambda_i)t_{i,j} + \delta_{ij}$$

amiből, bevezetve a $k_{i,j} = t_{i-1,j} - t_{i,j}$ jelölést,

$$k_{i+1,j} = \frac{\mu_i}{\lambda_i}k_{i,j} + \frac{\delta_{ij}}{\lambda_i}.$$

Ennek, a definíció szerint teljesülő $\sum_{l=1}^{n} k_{l,j} = 0$ egyenlőséget használva egyértelmű

megoldása van:

$$k_{i,j} = \begin{cases} -\rho_{i-1}\frac{1}{\lambda_j}\frac{1}{\rho_j}\frac{\sum_{l=j}^{n-1}\rho_l}{\sum_{l=0}^{n-1}\rho_l}, & i \le j \\ \\ -\rho_{i-1}\frac{1}{\lambda_j}\frac{1}{\rho_j}\frac{\sum_{l=j}^{n-1}\rho_l}{\sum_{l=0}^{n-1}\rho_l} + \frac{1}{\lambda_j}\frac{\rho_{i-1}}{\rho_j}, & i > j. \end{cases}$$

Mivel $t_{i,j} = -\sum_{l=1}^{i} k_{i,j}$, ezért

$$t_{i,j} = \begin{cases} \frac{1}{\lambda_j}\frac{1}{\rho_j}\sum_{l=0}^{i-1}\rho_l\frac{\sum_{m=j}^{n-1}\rho_m}{\sum_{m=0}^{n-1}\rho_m}, & i \le j \\ \\ \frac{1}{\lambda_j}\frac{1}{\rho_j}\left(\sum_{l=0}^{i-1}\rho_l\frac{\sum_{m=j}^{n-1}\rho_m}{\sum_{m=0}^{n-1}\rho_m} + \sum_{l=j+1}^{i-1}\rho_l\right), & i > j. \end{cases} \tag{2}$$

2.3. Születési folyamatok

Legyen $X(t)$ nemnegatív értékű, $[0,\infty)$ paramétertartományú Markov-folyamat, ami egy $\{\lambda_k\}$ pozitív számokból álló sorozatra teljesíti a következőket:

(i) $P(X(t+h) - X(t) = 1 | X(t) = k) = \lambda_k h + o(h)$ ha $h \downarrow 0$,

(ii) $P(X(t+h) - X(t) = 0 | X(t) = k) = 1 - \lambda_k h + o(h)$ ha $h \downarrow 0$,

(iii) $P(X(t+h) - X(t) < 0 | X(t) = k) = 0$.

Ekkor X-t tiszta születési folyamatnak nevezzük, a λ_k-kat születési intenzitásnak hívjuk. Ha $\lambda_k \equiv \lambda$, akkor Poisson-folyamatról beszélünk. A feltételekből adódik, hogy $X(t)$ bizonyos pontokban egyet "előre lép" az egész számokon, két ilyen lépés között pedig konstans. Ha ehelyett "hátra" lépéseket engedünk meg, akkor tiszta halálozási folyamatról beszélünk. T_j-vel jelöljük a j-dik és $(j+1)$-dik lépés között eltelt várakozási időt. Ekkor T_j exponenciális eloszlású λ_j paraméterrel, és T_j-k függetlenek.

3. A Wright-Fisher modell

3.1. Bevezetés

Tekintsünk egy $2N$ egyedből álló populációt. A generációkat a $t = 0, 1, 2, ...$ időpontokban fogjuk tekinteni. Feltesszük, hogy a generációk diszjunktak, és hogy a populáció mérete állandó. A $(t+1)$-dik generáció génjeit a t-dik generációból úgy kapjuk meg, hogy minden egyed a $(t+1)$-dik generációból egyenlő valószínűséggel választ magának szülőt a t-dik generációból, és a szülő génjét viszi tovább.

Megjegyezzük, hogy a populáció paritásának semmilyen matematikai jelentősége nincs, történeti okok miatt sok helyen használják a $2N$ méretű populációt, a továbbiakban mi is ezt tesszük.

A legegyszerűbb esetben mindössze két allélunk van, A_1 illetve A_2. Az A_1 típusú gének számát a t-dik generációban $X(t)$-vel jelölve tehát $X(t+1)$ binomiális eloszlású $2N$ renddel és $\frac{X(t)}{2N}$ paraméterrel. Másképp fogalmazva, az $X(t) = i$ feltétel mellett az $X(t+1) = j$ esemény p_{ij} valószínűsége

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}. \tag{3}$$

A fenti alakból nyilvánvaló, hogy $X(\cdot)$ a a $P = (p_{ij})$ átmenetmátrixszal Markov-láncot alkot, így $X(\cdot)$ viselkedése $X(0)$ és P ismeretében teljesen leírható.

Szintén könnyen látszik (3)-ből, hogy előbb-utóbb $X(\cdot)$ felveszi a 0 vagy $2N$ értéket, és ettől az időponttól $X(\cdot)$ konstans. A következőkben ezzel az elnyelődéssel kapcsolatos mennyiségeket vizsgálunk.

3.1. Állítás. *Annak a feltételes valószínűsége, hogy a két gén közül A_1 fixálódik,* $\frac{X(0)}{2N}$.

Bizonyítás. Jelölje π_i A_1 fixálódásának valószínűségét az $X(t) = i$ feltétel mellett. Az ehhez tartozó Markov-lánc elnyelődési valószínűségekre vonatkozó (1) egyenleteket

$$\hat{\pi}_i = \frac{i}{2N} \tag{4}$$

kielégíti:

$$\sum_{j=0}^{2N} p_{ij}\hat{\pi}_j = \sum_{j=0}^{2N} \frac{2N!}{j!\,(2N-j)!}\frac{j}{2N}\left(\frac{i}{2N}\right)^j\left(1-\frac{i}{2N}\right)^{2N-j} =$$

$$= \frac{i}{2N}\sum_{j=1}^{2N}\binom{2N-1}{j-1}\left(\frac{i}{2N}\right)^{j-1}\left(1-\frac{i}{2N}\right)^{2N-j} = \hat{\pi}_i,$$

így az egyértelműség miatt $\pi_i = \hat{\pi}_i$, és $i = X(0)$ helyettesítéssel kapjuk az állítást.

■

A következő vizsgált mennyiség a valamely gén fixálódásáig eltelt idő várható értéke. Erre explicit formula nem ismert; egy, a későbbiekben is használt közelítő módszert alkalmazunk. Vezessük be a következő jelöléseket: $\frac{i}{2N} = x$, $\frac{j}{2N} = x + \Delta$, továbbá jelölje $t(x)$ a fenti várható értéket, ha a A_1 jelenlegi relatív gyakorisága x. Ha feltesszük, hogy t kétszer folytonosan differenciálható függvénye a folytonos x argumentumnak, akkor (3)-t használva Taylor-sorfejtéssel a következőt kapjuk:

$$t(x) = 1 + \sum_{\delta} P(\Delta = \delta)t(x+\delta) = 1 + E(t(x+\Delta)) \approx$$

$$\approx 1 + t(x) + E(\Delta)t'(x) + \frac{1}{2}E(\Delta^2)t''(x). \tag{5}$$

(3)-ból tudjuk, hogy feltéve, hogy A_1 jelenlegi gyakorisága x, $2N(x+\Delta)$ eloszlása $Binom(2N, x)$, így

$$E(\Delta) = 0, \text{ illetve } E(\Delta^2) = \frac{x(1-x)}{2N}.$$

Tehát a (5) közelítésből a következő differenciálegyenlet adódik:

$$\frac{x(1-x)}{4N}t''(x) \approx -1.$$

Ennek a természetes $t(0) = t(1) = 0$ peremfeltételek mellett egyértelmű megoldása van, így

$$t(x) \approx -4N(x\log(x) + (1-x)\log(1-x)). \tag{6}$$

Fix x esetén az elnyelődésig eltelt idő N-ben lineáris. Ha viszont egy tiszta A_2 populációba bekerül egy A_1 gén, azaz $x = \frac{1}{2N}$, akkor igen gyorsan, $O(\log N)$ idő alatt ki is hal.

3.2. Egyirányú mutáció

Még mindig csak a két allélos esetre szorítkozva, most tegyük fel, hogy az A_1 allél u valószínűséggel A_2-be mutálódik, azaz az új generáció egy tagja u valószínűséggel akkor is A_2 lesz, ha az előző generációból A_1 szülőt választott. A később többször előforduló θ mennyiséget a Wright-Fisher modellben $\theta = 4Nu$-ként definiáljuk. A (3)-nak megfelelő átmeneti valószínűségek így a következők lesznek:

$$p_{ij} = \binom{2N}{j} (\psi_i)^j (1-\psi_i)^{2N-j} \tag{7}$$

ahol $\psi_i = (1-u)\frac{i}{2N}$. Az A_1 gén előbb-utóbb eltűnik, ennek a bekövetkezéséig eltelt időt probáljuk becsülni. Jelölje ismét $t(x)$ az eltűnésig eltelt idő várható értékét az $\frac{X(0)}{2N} = x$ feltétel mellett. (5)-t itt is változtatás nélkül felírhatjuk, azonban az új modellben $E(\Delta) = -ux$, illetve $E(\Delta^2) = \frac{x(1-x)}{2N} + O(\frac{1}{N^2})$, így a kapott differenciálegyenlet a következő:

$$-uxt'(x) + \frac{x(1-x)}{4N} t''(x) = -1.$$

Ez t'-re elsőrendű lineáris differenciálegyenlet, így könnyen megoldható. A $t(0) = 0$, illetve $\lim_{x\to 1} t(x) < \infty$ peremfeltételek melletti megoldása

$$t(x) = \int_0^1 \bar{t}(s,x)ds \tag{8}$$

ahol $\theta = 4Nu(\neq 1)$, $K = 1-(1-x)^{1-\theta}$ jelölésekkel

$$\bar{t}(s,x) = \begin{cases} 4Ns^{-1}(1-\theta)^{-1}((1-s)^{\theta-1}-1), & 0 < s \leq x \\ \\ 4NKs^{-1}(1-\theta)^{-1}(1-s)^{\theta-1}, & x \leq s \leq 1 \end{cases}$$

Mivel a differenciálegyenletet itt is a Taylor-sor első három tagjával való közelítéssel

kaptuk, (8) is valójában csak közelítés.

Későbbi hivatkozásnak megvizsgáljuk az $x = \frac{1}{2N}$ speciális esetet. További közelítéseket alkalmazva, $K \approx \frac{1-\theta}{2N}$ így a $t(s, \frac{1}{2N})$ függvény a következőképpen alakul:

$$\bar{t}\left(s, \frac{1}{2N}\right) \approx \begin{cases} 4N, & 0 < s \leq \frac{1}{2N} \\ \\ 2s^{-1}(1-s)^{\theta-1}, & \frac{1}{2N} \leq s \leq 1 \end{cases},$$

így

$$t\left(\frac{1}{2N}\right) \approx 2\left(1 + \int_{\frac{1}{2N}}^{1} s^{-1}(1-s)^{\theta-1} ds\right). \tag{9}$$

3.3. Kétirányú mutáció

Most tegyük fel, hogy A_2 is mutálódik A_1-be, (7) előtt definiált értelemben, v valószínűséggel. Ekkor (7)-ben ψ_i-t így értelmezzük:

$$\psi_i = \frac{(1-u)i + v(2N-i)}{2N}. \tag{10}$$

Ebben a modellben létezik az A_1 allélok számára stacionárius eloszlás, jelölje ezt $\Phi = (\Phi_0, \Phi_1, \ldots, \Phi_{2N})$, ahol Φ_i annak a stacionárius valószínűsége, hogy i darab A_1 gén van. Ez tehát teljesíti a $\Phi = \Phi P$ egyenletet a (7) illetve (10) egyenletek által adott P átmenetmátrixra. A stacionárius eloszlás μ várható értékére tehát

$$\mu = \Phi\xi = \Phi P\xi$$

ahol $\xi = (0,1,2,\ldots,2N)$. A $P\xi$ vektor i-dik koordinátája

$$\sum_{j=0}^{2N} j \binom{2N}{j} \psi_i^j (1-\psi_i)^{2N-j} = E(Binom(2N, \psi_i)) = 2N\psi_i = (1-u)i + v(2N-i).$$

Ezt használva

$$\mu = \Phi P \xi = \sum_{i=0}^{2N} \Phi_i((1-u)i + v(2N-i)) = (1-u)\sum_{i=0}^{2N} \Phi_i i + v\sum_{i=0}^{2N} \Phi_i(2N-i) =$$

$$= (1-u)\mu + v(2N-\mu)$$

ahonnan

$$\mu = \frac{2Nv}{u+v}.$$

Ezzel a módszerrel a magasabb momentumok is kiszámolhatóak.

3.2. Állítás. *Legyen $u = v$. Ekkor stacionárius eloszlás esetén annak a valószínűsége, hogy két véletlen választott gén azonos típusú, $\frac{1+2u(1-u)(2N-2)}{1+4u(1-u)(2N-1)}$*

Bizonyítás. A kerestt F_2 valószínűseg nyilván ugyanaz két egymást követő generációban. Két véletlen választott gén azonos szülőtől származik $\frac{1}{2N}$ valószínűséggel, különbözőtől $1-\frac{1}{2N}$ valószínűséggel, akik viszont azonos típúsúak F_2 valószínűséggel. A két gén közül 0, 1, vagy 2 mutálódott, ezek valószínűségei $(1-u)^2$, $2u(1-u)$, illetve u^2. Így tehát a következő egyenletet kapjuk F_2-re:

$$F_2 = (u^2 + (1-u)^2)\left(\frac{1}{2N} + F_2\left(1-\frac{1}{2N}\right)\right) + 2u(1-u)(1-F_2)\left(1-\frac{1}{2N}\right)$$

ahonnan átrendezéssel adódik az állítás. ■

A most kiszámolt mennyiség valamilyen értelemben a populáció homogenitását jellemzi. Hasonlóan kaphatjuk annak az F_i valószínűségét, hogy i kiválasztott gén azonos típusú. A fenti alakból adódik az $F_2 \approx \frac{1+\theta}{1+2\theta}$ közelítés is.

Az (3) által leírt modell könnyen kiterjeszthető $M > 2$ allélra. Ekkor $X_i(t)$-vel jelölve az A_i típusú allélok számát a t időpontban, a populációt az $\underline{X} = (X_1, X_2, \ldots, X_M)$ vektorral tudjuk leírni. Most is feltesszük, hogy a populáció állandó méretű, azaz minden t-re $X_1(t) + X_2(t) + \ldots + X_M(t) = 2N$. A (3)-nak megfelelő valószínűségek itt

$$P(\underline{X}(t+1) = \underline{k} \,|\underline{X}(t) = \underline{l}) = \frac{2N!}{k_1! \ldots k_m!}\left(\frac{l_1}{2N}\right)^{k_1} \ldots \left(\frac{l_M}{2N}\right)^{k_M}. \qquad (11)$$

A kétallélos eset néhány tulajdonsága csekély változtatással átvihető a többallélos modellre. Az A_i allél fixálódásának valószínűsége például itt is $\frac{X_i(0)}{2N}$ lesz: csoportosítsuk az allélokat az A_i és a nem-A_i osztályokra, és alkalmazzuk a 3.1. Állítást.

Mutációt bevezetve itt is létezik stacionárius eloszlás, és a 3.2. Állítás gondolatmenetét alkalmazva itt is kaphatunk formulát annak a valószínűségere, hogy két véletlen választott gén azonos típusú. Ha a mutáció teljesen szimmetrikus, azaz minden allél u valószínűséggel mutálódik, és minden más allélba egyenlő eséllyel, akkor a következő azonosságot írhatjuk fel:

$$F_2 = \left(\frac{1}{2N} + F_2\left(1 - \frac{1}{2N}\right)\right)\left((1-u)^2 + u^2\frac{1}{M-1}\right) +$$

$$+ \left(1 - \frac{1}{2N}\right)(1-F_2)\left(2u(1-u)\frac{1}{M-1} + u^2\left(1 - \frac{1}{M-1}\right)\frac{1}{M-1}\right)$$

amiből

$$F_2 \approx \frac{M-1+\theta}{M-1+M\theta}, \tag{12}$$

összhangban a két allélos esetben kapottakkal.

3.4. Végtelen allél

Az M-allélos modellt nem túl nagy M-re olyan esetben használhatjuk, ha a vizsgált géneket az általuk meghatározott tulajdonság szerint csoportosítjuk. Ilyen csoportosítás lehet például az ABO vércsoportért felelős géneké, $M = 3$-ra. Ha csoportosítás nélkül, minden allélt meg szeretnénk különböztetni, akkor egy 3000 nukleotidból álló génnek 4^{3000} különböző változata lehet, ezt már gyakorlati szempontból tekinthetjük végtelennek.

Nyilván a végtelen sok féle allélnak csak akkor lehet szerepe, ha megengedünk mutációt. A mutációról azonban itt egy új tulajdonságot is felteszünk, nevezetesen, hogy minden mutációnál új, eddig elő nem fordult allél keletkezik. A mutáció valószínűsége legyen egységesen u, ekkor az átmeneti valószínűségek a következők

lesznek:

$$P(X_i(t+1)=k_i \quad i=0,1,2,...|X_i(t)=l_i \quad i=1,2,...)=\frac{(2N)!}{\prod_{i=0}^{\infty}k_i!}\prod_{i=0}^{\infty}\pi_i^{k_i} \qquad (13)$$

ahol $X_i(t)$ jelöli a t generációban az A_i allélok számát ($i=1,2,...$), $X_0(t+1)$ pedig az új generációban keletkezett – és mind különböző – mutáns allélokét, és $\pi_0=u$ illetve $\pi_i=(1-u)\frac{l_i}{2N} \quad i=1,2,....$

Stacionárius eloszlásról most az eddigi értelemben nem beszélhetünk, hiszen minden allél előbb-utóbb eltűnik a populációból. Az egyes generációkban ezért az allélok száma helyett tekintsük az allélok által meghatározott partícióját $2N$-nek. (13) tekinthető úgy is, mint egy Markov lánc a lehetséges partíciókon. Bár a konkrét valószínűségek igen bonyolultak lesznek, az irreducibilitás és a véges állapottér miatt létezik stacionárius eloszlás. Ezzel kapcsolatosan vizsgálunk meg néhány mennyiséget.

Keressük meg először, hogy átlagosan hány allél lesz jelen a poluációban. Bármely A_i allélt kiválasztva, A_i valamikor belép a populációba $\frac{1}{2N}$ relatív gyakorisággal, majd egy idő után eltűnik, eközben a gyakorisága (7) szerint változik. A populációban töltött idő várható értékét jelöljük $E(T)$-vel ((8) jelölésével $E(T)=t(\frac{1}{2N})$).

3.3. Állítás. *Stacionárius esetben, ha $E(K_{2N})$ jelöli az egész populációban jelen levő allélok számának várható értékét, akkor $2Nu=\frac{E(K_{2N})}{E(T)}$.*

Bizonyítás. Minden generációban várhatóan $2Nu$ új allél jön létre, tehát r egymást követő generációban összesen $r2Nu+O(1)$ allél van jelen. Másképp számolva, az egyes generációkban jelen levő allélok számát összeadva, $rE(K_{2N})$-et kapunk, de ekkor minden allélt átlagosan $E(T)$ generációban is megszámoltunk. Tehát

$$r2Nu+O(1)=r\frac{E(K_{2N})}{E(T)}$$

és $r\to\infty$ adja az állítást. ■

Felhasználva az állítást és a (9) közelítést,

$$E(K_{2N})=2NuE(T)=2Nut\left(\frac{1}{2N}\right)\approx\theta+\int_{\frac{1}{2N}}^{1}\theta x^{-1}(1-x)^{\theta-1}dx.$$

A 3.2. Állításhoz hasonlóan kiszámolhatjuk most is az $F_2, F_3, ...$ mennyiségeket, ehelyett azonban most egy jóval általánosabb formulát bizonyítunk. Ha az F_n mennyiségre úgy tekintünk, mint annak a valószínűségére, hogy egy véletlen kiválasztott n elemű mintán az allélok az $\{n\}$ partíciót határozzak meg, akkor megkérdezhetjük azt is, hogy tetszőleges π partíciónak mekkora a valószínűsége. Legyen tehát a minta mérete elhanyagolható a populációhoz képest és vezessük be a következő jelöléseket: B_i legyen azoknak az alléloknak a száma amiknek pontosan i reprezentánsa van a mintában $i = 1,2, ..., n$, $\mathbf{B} = (B_1, B_2, ...B_n)$.

3.4. Tétel. (Ewens Sampling Formula(ESF)) *Minden $b = (b_1, b_2, ..., b_n)$ vektorra, amire $\sum_{i=1}^{n} ib_i = n$,*

$$P(\boldsymbol{B} = b) \approx \frac{n!}{S_n(\theta)} \prod_{i=1}^{n} \frac{\theta^{b_i}}{i^{b_i} b_i!} \tag{14}$$

ahol $S_n(\theta) = \theta(\theta+1)(\theta+2)\dots(\theta+n-1)$.

Bizonyítás. Az approximációt olyan értelemben fogjuk bizonyítani, hogy $N \to \infty$, $u \to \infty$ esetén, de fix $\theta = 4Nu$ és n mellett a két oldal különbsége $O\left(\frac{1}{N^2}\right)$. Legyen a kiválasztott mintában n_1 az egyik fajta, n_2 egy másik fajta, és így tovább, n_k egy k-dik fajta allélból. b-nek megfelel egy $\{n_1, ..., n_k\}$ rendezetlen halmaz, így a keresett valószínűséget írhatjuk $P(n_1, n_2, ..., n_k)$ alakban is. Az egyszerűség kedvéért feltehetjük, hogy $n_1 \geq n_2 \geq ... \geq n_k$. Definíció szerint b_1 az egy reprezentánssal rendelkező allélok száma. b_1 és n szerinti indukcióval bizonyítunk: először belátjuk $b_1 = 0$ esetre, majd feltéve, hogy $n \leq r$ és $b_1 \leq m$, illetve $n = r+1$ és $b_1 < m$ esetekre tudjuk (14)-t, belátjuk, hogy $n = r+1$ és $b_1 = m$ esetben is igaz. A $b_1 = = 0$ eset az indukciós lépeshez hasonló okoskodással bizonyítható, ezért csak utóbbit részletezzük.

Próbáljuk tehát közelíteni a $P(n_1, ..., n_k)$ mennyiséget $n = r+1$ és $b_1 = m \geq \geq 1$ esetén. Legyen q_i annak a valószínűsége, hogy az $r+1$ gén pontosan i szülőtől származik. Az $x^{[i]} = x(x-1)\dots(x-i+1)$ jelöléssel

$$q_{r+1} = \frac{(2N)^{[r+1]}}{(2N)^{r+1}} = 1 - \frac{r(r+1)}{4N} + O\left(\frac{1}{N^2}\right),$$

$$q_r = \binom{r+1}{2} \frac{(2N)^{[r]}}{(2N)^{r+1}} = \frac{r(r+1)}{4N} + O\left(\frac{1}{N^2}\right),$$

innen pedig nyilván $q_1 + q_2 + ... + q_{r-1} = O\left(\frac{1}{N^2}\right)$. Q_i jelölje az $\{n_1, ..., n_k\}$ konfiguráció megfigyelésének a valószínűséget a (pontosan i szülő) feltétel mellett. Ekkor $P(n_1, ..., n_k) = q_{r+1}Q_{r+1} + q_rQ_r + O\left(\frac{1}{N^2}\right)$, így már csak a Q_{r+1} és Q_r mennyiségek kiszámolása van hátra. Ezekre a következőket kapjuk:

$$Q_{r+1} = P(n_1, ..., n_k)(1-u)^{r+1-m} + \sum_{j \le k-1} \frac{1}{b_{n_j}} P(n_1, ..., n_{j-1}, n_j + 1, n_{j+1}, ..., n_{k-1}) \times$$

$$\times u(1-u)^{r+1-m}(n_j+1)(b_{n_j+1}+1) + O(u^2).$$

Az első tag felel meg annak, hogy a szülők az előző generációban az $\{n_1, ..., n_k\}$ konfigurációt alkották, és a több reprezentánssal rendelkező allélok nem mutálódtak. A második tagban az $\frac{1}{b_{n_j}}$ tényező azért szükséges, mert az $\{n_1, ..., n_{j-1}, n_j+1, n_{j+1}, ..., n_{k-1}\}$ szülői konfiguráció b_{n_j}-szer jelenik meg az összegben; a $u(1-u)^{r+1-m}(n_j+1)(b_{n_j+1}+1)$ tényező pedig annak a valószínűsége, hogy egy olyan allél utódja mutálódik, aminek a szülői konfigurációban $n_j + 1$ reprezentánsa van, míg a többi, több reprezentánssal rendelkező allél nem. Hasonló gondolatmenettel kapjuk:

$$Q_r = \sum_j \frac{1}{b_{n_j}} P(n_1, ..., n_{j-1}, n_j - 1, n_{j+1}, ..., n_k) \frac{(n_j - 1)(b_{n_j-1} + 1)}{r} (1-u)^{r-m} + O(u).$$

A $P(n_1, ..., n_k) = q_{r+1}Q_{r+1} + q_rQ_r + O\left(\frac{1}{N^2}\right)$ egyenletbe tehát behelyettesítve a kapott eredményeket, $N \to \infty$, $u \to 0$, $4Nu \equiv \theta$ esetén átrendezés után a következőt kapjuk:

$$(r(r+1) + \theta(r+1-m))P(n_1, ..., n_k) = \theta \sum_j P(n_1, ...n_{j-1}, n_j + 1, n_{j+1}, ..., n_{k-1}) \times$$

$$\times \frac{(n_j+1)(b_{n_j+1}+1)}{b_{n_j}} + r(r+1) \sum_j P(n_1, ..., n_{j-1}, n_j - 1, n_{j+1}, ..., n_k) \frac{(n_j-1)(b_{n_j-1}+1)}{r b_{n_j}}.$$

A jobboldalon álló valószínűségekre alkalmazhatjuk az indukciós feltevést, ennek megfelelően (14)-t helyettesítve $P(n_1, ..., n_k)$-ra is megkapjuk a kívánt formulát. ■

3.5. Következmény. *$K_n = \sum_{i=1}^n B_i$ jelölje a mintában előforduló különböző allélok*

számát. Jelölje továbbá $|S_n^k|$ *a* θ^k *együtthatóját* $S_n(\theta)$*-ban. Ekkor*

$$P(K_n = k) \approx \frac{|S_n^k|\theta^k}{S_n(\theta)}.$$

(A közelítést most úgy értjük, hogy a ESF-ben egyenlőséget feltételezve itt is egyenlőséget kapunk.)

Bizonyítás. A ESF-ből kapott valószínűségeket kell összegeznünk olyan b vektorokra, amikre $\sum_{i=1}^n b_i = k$. Ez nyilván

$$\sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \frac{n!\,\theta^{\sum b_j}}{S_n(\theta)\prod_{j=1}^n j^{b_j}b_j!} = \frac{(n-1)!\,\theta^k}{S_n(\theta)} \sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \frac{n}{\prod_{j=1}^n j^{b_j}b_j!}.$$

A $\sum jb_j = n$ egyenlőséget felhasználva az összegzés így alakítható:

$$\sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \frac{\sum_{l=1}^n lb_l}{\prod_{j=1}^n j^{b_j}b_j!} = \sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \sum_{\substack{l=1..n,\\ b_l\neq 0}} \frac{1}{\prod_{j=1}^n j^{b_j^{(l)}}b_j^{(l)}!}$$

ahol a $b_j^{(l)}$ számokat így definiáljuk:

$$b_j^{(l)} = \begin{cases} b_j, & j \neq l \\ b_j - 1, & j = l. \end{cases}$$

Ezekre egyrészt $\sum_{j=1}^n b_j^{(l)} = k-1$, másrészt $\sum_{j=1}^n jb_j^{(l)} = n-l$, így

$$\sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \sum_{\substack{l=1..n,\\ b_l\neq 0}} \frac{1}{\prod_{j=1}^n j^{b_j^{(l)}}b_j^{(l)}!} = \sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \sum_{\substack{l=1..n,\\ b_l\neq 0}} \frac{1}{(n-l)} \frac{\sum_{k=1}^n kb_k^{(l)}}{\prod_{j=1}^n j^{b_j^{(l)}}b_j^{(l)}!} =$$

$$\sum_{\substack{b:\,\sum b_j=k,\\ \sum jb_j=n}} \sum_{\substack{l=1..n,\\ b_l\neq 0}} \sum_{\substack{k=1..n,\\ b_k^{(l)}\neq 0}} \frac{1}{(n-l)} \frac{1}{\prod_{j=1}^n j^{b_j^{(l,k)}}b_j^{(l,k)}!} =$$

$$\sum_{\substack{b:\,\sum b_j=k,\\ \sum j b_j=n}} \sum_{\substack{l=1..n,\\ b_l\neq 0}} \sum_{\substack{k=1..n,\\ b_k^{(l)}\neq 0}} \frac{1}{(n-l)} \frac{1}{(n-l-k)} \frac{\sum_{m=1}^{n} m b_m^{(l,k)}}{\prod_{j=1}^{n} j^{b_j^{(l,k)}} b_j^{(l,k)}!} = \ldots$$

ahol $b_j^{(l,k)}$-t analóg módon definiáljuk, és amikre tehát $\sum_{j=1}^{n} b_j^{(l,k)} = k-2$, illetve $\sum_{j=1}^{n} j b_j^{(l,k)} = n-l-k$. Az eljárást folytatva $(k-1)$ lépés után $(k-1)$ darab különböző, n-nél kisebb pozitív egész szám reciprokának a szorzata jelenik meg, a nevezőben levő produktum értéke pedig 1-é egyszerűsödik. Tehát

$$P(K_n = k) \approx \frac{\theta^k (n-1)!}{S_n(\theta)} \sum_{\substack{1\le m_i\le n-1\\ m_i\neq m_j\ i\neq j}} \frac{1}{\prod_{i=1}^{k-1} m_i} = \frac{\theta^k}{S_n(\theta)} \sum_{\substack{1\le l_i\le n-1\\ l_i\neq l_j\ i\neq j}} \prod_{i=1}^{n-k} l_i.$$

A jobboldalon álló összeg viszont épp $|S_n^k|$, így az állítást beláttuk. ■

K_n eloszlásának ismeretében néhány további mennyiség könnyen kiszámolható.

3.6. Következmény. *1.* $F_n = P(K_n = 1) \approx \frac{(n-1)!}{(\theta+1)(\theta+2)\ldots(\theta+n-1)}$,

2. $E(K_n) \approx \theta \sum_{j=0}^{n-1} \frac{1}{\theta+j}$,

3. $D^2(K_n) \approx \theta \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}$.

Speciálisan $F_2 \approx \frac{1}{1+\theta}$, amit az M allél esetén kapott (12) határértékeként is megkapunk, ha $M \to \infty$.

3.5. ESF alternatív levezetése

Megmutatható, hogy (13) esetén, ha a populáció méretéhez képest kis méretű mintát vizsgálunk, akkor sorban húzva a minta elemeit, annak a valószínűsége hogy a $(j+1)$-dik elem az eddigi j mindegyikétől különböző típusú, közelítőleg $\frac{\theta}{\theta+j}$. Könnyen látható továbbá, hogy annak a valószínűsége, hogy a $(j+1)$-edik elem olyan típusú lesz, aminek a mintában már m darab reprezentánsa van, $\frac{m}{\theta+j}$. A minta ezen tulajdonságát egy urna-modellel írhatjuk le.

Tekintsünk egy urnát, amiben egy $\theta > 0$ súlyú fekete, és több más színű, egységnyi súlyú golyó van. A j-dik lépésben a súlyokkal arányos valószínűséggel kihúzunk egy

golyót, és ha ez nem a fekete, akkor visszatesszük, és még egy ugyanolyan színű golyót teszünk az urnába. Ha fekete golyót húztunk, akkor a feketét és egy új, eddig nem használt színű golyót teszünk vissza. A (nem fekete) színeket a természetes számokkal jelöljuk. Az X_j valószínűségi változó jelölje a j-dik lépésben újonnan betett golyó színének számát. A legelső lépésben az urnában csak a fekete golyó van, így $X_1 = 1$, $X_2 = 1$ vagy 2, $X_3 = 1{,}2$ vagy 3, stb.

Legyen K az n-dik lépés után az urnában levő különböző, nem fekete színek száma. A fekete golyót a továbbiakban figyelmen kívül hagyjuk, a folyamatban csak az új színek generálása a szerepe. Az n-dik lépés után tehát n golyó van az urnában, az i-dik színből n_i darab ($i = 1{,}2, ..., K$). Tekintsük az $\{n_1, n_2, ..., n_K\}$ rendezetlen halmazt, ez a 3.4. Tételhez hasonlóan meghatározza n egy $\mathbf{B}_n$ véletlen partícióját: jelölje B_i, hogy hányszor fordul elő az i szám az $\{n_1, n_2, ..., n_K\}$ halmazban.

3.7. Tétel. *$\boldsymbol{B}_n$ Markov-láncot alkot, és a marginális eloszlása*

$$P(\boldsymbol{B}_n = b) = \frac{n!}{S_n(\theta)} \prod_{i=1}^{n} \frac{\theta^{b_i}}{i^{b_i} b_i!} \tag{15}$$

ha $b = (b_1, b_2, ..., b_n)$-re teljesül $\sum_{i=1}^{n} i b_i = n$.

Bizonyítás. Rögzítsünk egy $\{n_1, ..., n_K\}$ halmaznak megfelelő $b = (b_1, b_2, ..., b_n)$ partíciót, és vizsgáljuk meg egy, b-t eredményező $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ húzássorozat valószínűségét. Ehhez a θ súlyú fekete golyót K-szor kell kihúznunk, majd minden i-re az i-dik színből húznunk kell még $(n_i - 1)$-szer. Az urnában levő golyók összsúlyának szorzata az n lépés során $S_n(\theta)$, ezekből

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \frac{\theta^K \prod_{i=1}^{K} (n_i - 1)!}{S_n(\theta)}. \tag{16}$$

Most számoljuk meg a b partíciót eredményező húzások számát. Ez könnyen láthatóan megegyzik n darab, $1{,}2, ..., K$ színekre színezett tárgy olyan különböző permutációival, amik eleget tesznek a következő két feltételnek:

(i) Az első 1-es színű tárgy megelőzi az első 2-es színű tárgyat, ami viszont megelőzi az első 3-as színűt, stb.

(ii) A különböző színű tárgyak száma nincs meghatározva, csak annyit tudunk, hogy valamely színből n_1 darab van, egy másikból n_2 darab, stb.

Az $n_1, n_2, ..., n_K$ számokat a K szín között $\frac{K!}{\prod_{i=1}^{n} b_i!}$ féleképpen tudjuk elosztani. Minden ilyen elosztásra az n tárgyat $\frac{n!}{\prod_{i=1}^{K} n_i!}$ féleképpen tudjuk megszínezni. Az (i) feltételt tehát egyelőre figyelmen kívül hagyva, az (ii) feltételt

$$\frac{K!\, n!}{\prod_{i=1}^{n} b_i! \prod_{i=1}^{K} n_i!} \tag{17}$$

permutáció elégíti ki.

Osszuk most ezeket diszjunkt osztályokba aszerint, hogy az első tárgyak az 1,2, ..., K színekből milyen sorrendben fordulnak elő. $K!$ darab osztályt kapunk, szimmetria okokból mindegyik egyenlő elemszámú, és pontosan egy osztály elemei elégítik ki (i)-t. A keresett valószínűség tehát, (16) és (17) összeszorzása és $K!$-al való osztás után

$$\frac{n!}{S_n(\theta)} \frac{\theta^K}{\prod_{i=1}^{n} b_i! \prod_{i=1}^{K} n_i}.$$

A $\theta^K = \prod_{i=1}^{n} \theta^{b_i}$, illetve $\prod_{i=1}^{K} n_i = \prod_{i=1}^{n} i^{b_i}$ azonosságok alkalmazásával adódik (15). ■

4. A Moran-modell

4.1. Bevezetés

A Moran-modellben az egyik jelentős különbséget az fogja jelenteni, hogy nem tesszük fel a generációk diszjunktságát. Sőt, egy időpontban csak egy születést és halálozást engedünk meg. A megfelelő Markov lánc így több szempontból könnyebben kezelhető lesz, és így számos mennyiséget pontosan, közelítések nélkül tudunk meghatározni.

Ismét egy $2N$ egyedből álló állandó méretű populációt tekintünk. A $t = 1,2,...$ időpontokban véletlenszerűen választunk egy egyedet, aki szaporodik, majd ismét véletlenszerűen egy újabb egyedet, aki meghal. Ez utóbbi nem lehet az új egyed, de lehet az előzőleg választott szülő.

A kétallélos esetet vizsgálva, jelölje $X(t)$ az A_1 allélok számát a t időpontban. Ha $X(t) = i$, akkor $X(t+1)$ lehetséges értékei $i-1, i, i+1$, és az átmeneti valószínűségek

$$p_{i,\,i-1} = p_{i,\,i+1} = \frac{i(2N-i)}{(2N)^2}$$

$$p_{i,\,i} = \frac{(i^2 + (2N-i)^2)}{(2N)^2}.$$

$X(t)$ átmenetmátrixa tehát kontinuáns, és $\rho_i = 1$, $i = 0,1,...,2N$.

A 2.3. Állítás alapján, ha az indulási időpontban i darab A_1 gén van, akkor A_1 fixálódásának valószínűsége $\frac{i}{2N}$.

Az elnyelődésig eltelt idő várható értékét, i darab A_1 génnel indulva (2) alapján kaphatjuk:

$$t_{i,j} = \begin{cases} 2N\frac{i}{j}, & i \leq j \\ 2N\frac{2N-i}{2N-j}, & i > j, \end{cases}$$

így

$$t_i = 2N\left((2N-i)\sum_{j=1}^{i}\frac{1}{2N-j} + i\sum_{j=i+1}^{2N-1}\frac{1}{j}\right).$$

A fenti két összeg két integrálközelítő összegként is felfogható. Ha tehát $x = \frac{i}{2N}$ -el

jelöljük A_1 relatív gyakoriságát kezdetben, akkor az egyik allél fixálódásáig szükséges idő várható értéke

$$t(x) \approx -(2N)^2(x \log x + (1-x)\log(1-x)),$$

ami (6)-tól csak egy N szorzóban különbözik. Mivel a Moran-modellben $2N$ új egyed születéséhez $2N$, míg a Wright-Fisher modellben 1 idő szükséges, egy $2N$ szorzót intuitíven is jogosnak érezhetünk.

4.2. Mutációk

Legyen az $A_1 \to A_2$ mutáció valószínűsége u, az $A_2 \to A_1$-é pedig v. A modell a következőképpen módosul:

$$\begin{aligned} p_{i,i-1} = \mu_i &= \frac{i(2N-i)(1-v)+ui^2}{(2N)^2} \\ p_{i,i+1} = \lambda_i &= \frac{(i(2N-i)(1-u)+v(2N-i)^2)}{(2N)^2} \\ p_{i,i} &= 1 - p_{i,i-1} - p_{i,i+1}. \end{aligned} \tag{18}$$

A stacionárius eloszlást a 2.4. Állítás alapján kaphatjuk meg. Átalakítás után a

$$\phi_j = \phi_0 \frac{(2N)!\,\Gamma(j+A)\Gamma(B-j)}{j!\,(2N-j)!\,\Gamma(A)\Gamma(B)} \tag{19}$$

alakot kapjuk, ahol $A = \frac{2Nv}{(1-u-v)}$, $B = \frac{2N(1-v)}{(1-u-v)}$. A binomiális együtthatók általánosabb $\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}$ definícióját használva tehát

$$\phi_j = \phi_0 \frac{\binom{A-1+j}{A-1}\binom{B-1-j}{B-2N-1}}{\binom{B-1}{2N}}.$$

Felhasználva a $\sum_{j=0}^{2N} \phi_j = 1$ feltételt illetve a binomiális azonosságokat,

$$\phi_0 = \frac{\binom{B-1}{2N}}{\binom{A+B-1}{A+B-2N-1}} = \frac{\Gamma(B)\Gamma(A+B-2N)}{\Gamma(A+B)\Gamma(B-2N)}.$$

Nagy a és a-hoz képest kis b esetén $\frac{\Gamma(a+b)}{\Gamma(a)} \sim a^b$. Ha tehát most N-et és j-t növeljük, u-t és v-t pedig csökkentjük úgy, hogy az $x = \frac{j}{2N}$, $\alpha = 2Nu$, $\beta = 2Nv$ mennyiségek fixek maradjanak, akkor elég nagy N-re

$$\phi_j \approx \frac{1}{2N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\beta-1}(1-x)^{\alpha-1}.$$

Ez a közelítés, bár a (19) alakkal szemben nem pontos, de nyilván lényegesen jobban kezelhető.

4.3. Végtelen allél

A Wright-Fisher modellhez hasonlóan vezethetünk be itt is végtelen sok allélt: az újonnan született gén u valószínűséggel mutálódjon, és minden mutációnál új allél jöjjön létre. Stacionárius eloszlásrol ismét a 3.4. Tétel előtti értelemben beszélhetünk. Számos, a Wright-Fisher modellben megismert eredmény analógja levezethető. Be fogjuk látni a ESF-t, ami most pontos valószínűségeket fog adni, ráadásul akármekkora mintára. Speciálisan az egész populációra fogjuk tudni a stacionárius eloszlást, ennek a bizonyításával kezdjük.

4.1. Tétel. *Jelölje B_i az i darab reprezentánssal rendelkező gének számát a populációban, $\boldsymbol{B} = (B_1, B_2, ..., B_{2N})$. Ekkor stacionárius esetben minden $b = (b_1, b_2, ..., b_{2N})$ vektorra, amire $\sum_{j=1}^{2N} j b_j = 2N$,*

$$P(\boldsymbol{B} = b) = \frac{2N!}{S_{2N}(\theta)} \prod_{i=1}^{2N} \frac{\theta^{b_i}}{i^{b_i} b_i!}, \tag{20}$$

ahol $\theta = \frac{2Nu}{1-u}$.

Bizonyítás. Jelöljük a definiált eloszlást $\pi(b)$-vel. Megmutatjuk, hogy π kielégíti a $\pi(b)p_{bb'} = \pi(b')p_{b'b}$ egyenleteket minden b, b'-re, azaz a megfelelő Markov-lánc megfordítható és valóban stacionárius eloszlást kapunk. Jelölje továbbá L_i az i reprezentánnsal rendelkező alléllal rendelkező egyedek halmazát, nyilván $|L_i| = ib_i$.

Nézzük meg, hogy ha a mostani generációban az allélok által meghatározott konfiguráció b, akkor a következő generációban milyen b' konfiguráció jöhet szóba. Mivel

a $b = b'$ eset triviális, ezért a továbbiakban feltesszük hogy különböző típusú gén szaporodott és halt meg, vagy pedig mutáció történt. Ha történt mutáció, akkor a szaporodó gén típúsa lényegtelen, a meghaló gén pedig legyen L_j-beli. Ekkor $b'_m = = b_m + \delta_{1m} + \delta_{j-1,m} - \delta_{jm}$. Ha nem volt mutáció, és egy L_i-beli gén szaporodott, továbbá egy (előbbitől eltérő típusú) L_j-beli halt meg, akkor $b'_m = b_m + \delta_{i+1,m} - - \delta_{im} + \delta_{j-1,m} - \delta_{jm}$. Ebből egyrész leolvasható, hogy $p_{bb'} \neq 0 \Leftrightarrow p_{b'b} \neq 0$, másrészt a pozitív valószínűséggel szóba jövő esetek:

1. $b' = (b_1 + 1, ..., b_{j-1} + 1, b_j - 1, ..., b_{2N})$,
2. $b' = (b_1 + 2, b_2 - 1, ..., b_{2N})$,
3. $b' = (b_1, b_2, ..., b_{j-1} + 1, b_j - 1, ..., b_i - 1, b_{i+1} + 1, ..., b_{2N})$,
4. $b' = (b_1, ..., b_{i-1} + 1, b_i - 2, b_{i+1} + 1, ..., b_{2N})$,
5. $b' = (b_1, ..., b_{i-1} - 1, b_i + 2, b_{i+1} - 1, ..., b_{2N})$.

A bizonyítás mind az 5 esetre nagyon hasonló, ezért most csak az 1. esetet részletezzük. Ha $b' = (b_1 + 1, ..., b_{j-1} + 1, b_j - 1, ..., b_{2N})$, akkor egy mutációnak kellett történnie, továbban egy L_j-beli egyednek meghalnia. Fordítva, ha b'-ből indulunk ki, akkor ahhoz, hogy a b konfigurációt kapjuk, egy L_{j-1}-belinek kell szaporodnia és egy L_1-belinek meghalnia. Így

$$\pi(b)p_{bb'} = \frac{2N!\,\theta^{\sum_{i=1}^{2N} b_i}}{1^{b_1}2^{b_2}\cdots 2N^{b_{2N}} b_1!\,b_2!\cdots b_{2N}!\,S_{2N}(\theta)}\, u\frac{(1-u)jb_j}{2N} =$$

$$\frac{2N!\,\theta^{\sum_{i=1}^{2N} b_i}}{1^{b_1}\cdots j^{b_j-1}\cdots 2N^{b_{2N}} b_1!\cdots (b_j-1)!\cdots b_{2N}!\,S_{2N}(\theta)}\frac{\theta}{(2N)^2} =$$

$$= \frac{2N!\,\theta^{(\sum_{i=1}^{2N} b_i)+1}}{1^{b_1+1}\cdots (j-1)^{b_j-1+1} j^{b_j-1}\cdots 2N^{b_{2N}} (b_1+1)!\cdots (b_{j-1}+1)!\,(b_j-1)!\cdots b_{2N}!\,S_{2N}(\theta)}\times$$

$$\times\frac{b_1(j-1)(b_{j-1}+1)}{(2N)^2} = \pi(b')p_{b'b}$$

■

Most az ESF egy újabb tulajdonságát látjuk be: ha a teljes populáció (20) eloszlású, akkor minden véletlen választott minta eloszlása is ilyen típusú lesz.

4.2. Tétel. *Legyen $1 \leq n \leq 2N$ tetszőleges. Ekkor ha az n elemű mintában az i reprezentánssal rendelkező allélok számát B_i jelöli, $B^n = (B_1, B_2, ..., B_n)$, akkor minden olyan $b = (b_1, b_2, ..., b_n)$ vektorra, amire $\sum_{i=1}^{n} ib_i = n$ teljesül,*

$$P(B^n = b) = \frac{n!}{S_n(\theta)} \prod_{i=1}^{n} \frac{\theta^{b_i}}{i^{b_i} b_i!}.$$

Bizonyítás. Az $n = 2N$ esetet az előbbi tételben láttuk. Ebből megmutatjuk, hogy $n = 2N - 1$ esetben is igaz a formula, ezt a lépést ismételve látható minden n-re. Legyen tehát $b = (b_1, b_2..., b_{2N})$ olyan, hogy $\sum_{i=1}^{2N} ib_i = 2N - 1$. Alkalmazva a teljes valószínűség tételét és felhasználva (20)-t:

$$P(B^{2N-1} = b) = \sum_{b'} P(\mathbf{B} = b') P(B^{2N-1} = b | \mathbf{B} = b') =$$

$$= P(\mathbf{B} = (b_1 + 1, b_2, ..., b_{2N})) \frac{b_1 + 1}{2N} + P(\mathbf{B} = b_1 - 1, b_2 + 1, b_3, ..., b_{2N}) \frac{2(b_2 + 1)}{2N} + \cdots$$

$$\cdots + P(\mathbf{B} = (b_1, ..., b_{2N-1} - 1, b_{2N} + 1)) \frac{2N(b_2 N + 1)}{2N} =$$

$$= \frac{(2N-1)!\,\theta^{(\sum b_i)+1}}{1^{b_1} \cdots 2N^{b_{2N}} b_1! \cdots b_{2N}!\, S_{2N}(\theta)} + \sum_{b_j > 0} \frac{(2N-1)!\,\theta^{\sum b_i}}{1^{b_1} \cdots j^{b_j - 1} \cdots 2N^{b_{2N}} b_1! \cdots (b_j - 1)! \cdots b_{2N}!\, S_{2N}(\theta)} =$$

$$= \frac{(2N-1)!\,\theta^{\sum b_i}}{1_1^b \cdots 2N^{b_{2N}} b_1! \cdots b_{2N}!} \left(\frac{\theta + \sum_{b_j > 0} j b_j}{S_{2N}(\theta)} \right) = \frac{(2N-1)!\,\theta^{\sum b_i}}{1_1^b \cdots 2N^{b_{2N}} b_1! \cdots b_{2N}!\, S_{2N-1}(\theta)},$$

hiszen $\sum_{b_j > 0} j b_j = \sum_{j=1}^{2N} j b_j = 2N - 1$. Innen átrendezéssel kapjuk a tételbeli alakot. ■

A megfelelő következmények is érvényben maradnak, ismét pontos eredményeket adva, és ismét bármekkora mintára.

4.3. Következmény. *1.* $P(K_n = k) = \frac{|S_n^k| \theta^k}{S_n(\theta)}$,

2. $F_n = P(K_n = 1) = \frac{(n-1)!}{(\theta+1)(\theta+2)...(\theta+n-1)}$,

3. $E(K_n) = \theta \sum_{j=0}^{n-1} \frac{1}{\theta+j}$,

4. $D^2(K_n) = \theta \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}$.

5. A Cannings-modell

5.1. Sajátértékek

Röviden foglalkozunk az igen általános Cannings-modellel, aminek az előzőekben tárgyalt két modell speciális esete. A nagyfokú általánosság mellett egyszerű képletet kapunk a Markov-láncot meghatározó átmenetmátrix sajátértékeire. Csak a két allélos, mutáció nélküli esettel foglalkozunk.

Jelölje ismét $X(t)$ az A_1 allélok számát és egy adott generációban az i-edik egyed utódainak száma legyen Y_i. Ezek nemnegatív értékű valószínűségi változók, és teljesül $\sum_{i=1}^{2N} Y_i = 2N$. Ezen felül csupán annyit teszünk fel, hogy felcserélhetőek, azaz bármely $(Y_{n_1}, Y_{n_2}..., Y_{n_k})$ és $(Y_{m_1}, Y_{m_2}, ..., Y_{m_k})$ részhalmazokat kiválasztva, ezek együttes eloszlása megegyezzen (speciálisan minden Y_i azonos eloszlású). Ezek alapján $X(t+$ $+1) = \sum_{k=1}^{X(t)} Y_k$, azaz $P_{ij} = P(\sum_{k=1}^{i} Y_k = j)$ $(i, j = 0,1, ...,2N)$.

Az Y_l utód-változókat felfoghatjuk $Y_l = Y_l^0 + Y_l^1$ alakban is, ahol Y_l^0 $0-1$ értékű változó. Ekkor Y_l^0 jelöli azt, hogy az l-dik egyed életben maradt-e a következő generációban vagy sem, Y_l^1 pedig a tényleges utódok számát. Így a Cannings-modell valóban magába foglal egymást átfedő generációkat megengedő modelleket, többek között a Moran-modellt is.

5.1. Tétel. *Az fent definiált P mátrix sajátértékei:*

$$\lambda_0 = 1, \quad \lambda_j = E(\prod_{k=1}^{j} Y_k) \quad j = 1,2, ...,2N.$$

A felcserélhetőség miatt természetesen bármelyik j darab Y_k-t használhatjuk λ_j kifejezésében.

Bizonyítás. Először egy lemmát látunk be, aminek a tétel már egyszerű következménye lesz.

5.2. Lemma. *Legyen $\{X_t\}$ Markov-lánc, állapottere $(0,1, ..., n)$, átmenetmátrixa P. Ha az $X_t = i$ feltétel mellett X_{t+1} j-edik momentuma i-nek legfeljebb j-edfokú polinomja, azaz*

$$E(X_{t+1}^j | X_t = i) = \lambda_j i^j + \lambda_{j-1,j} i^{j-1} + \cdots + \lambda_{0,j},$$

minden $j = 0,1,...,n$ esetén, akkor a λ_j együtthatók a P mátrix sajátértékei.

Bizonyítás. A korábban bevezetett $x^{[i]}$ jelölést használva a fenti azonosság nyilvánvalóan ekvivalens átfogalmazása:

$$E(X_{t+1}^{[j]}|X_t = i) = \lambda_j i^j + \lambda'_{j-1,j} i^{j-1} + \cdots + \lambda'_{0,j}.$$

Vezessük be a $B = \{b_{ij}\} = \{\binom{i}{j}\}$ $(n+1) \times (n+1)$-es mátrixot, ekkor $B^{-1} = \{b^{ij}\} = \{(-1)^{i+j}\binom{i}{j}\}$. Az $A = B^{-1}PB$ mátrix sajátértekei megegyeznek P sajátértékeivel. Másrészt A-ról belátjuk, hogy felső háromszögmátrix, így sajátértékei az átlóban levő elemek, amik pedig épp λ_j-k lesznek.

$$a_{kj} = (B^{-1}(PB))_{kj} = \sum_{i=0}^{n} b^{ki}(PB)_{ij} = \sum_{i=0}^{n} b^{ki} \sum_{l=0}^{n} p_{il} \binom{l}{j} = \sum_{i=0}^{n} b^{ki} \frac{1}{j!} E(X_{t+1}^{[j]}|X_t = i) =$$

$$= (-1)^k \sum_{i=0}^{n} (-1)^i \binom{k}{i} \frac{1}{j!} E(X_{t+1}^{[j]}|X_t = i) = (-1)^k \sum_{i=0}^{k} (-1)^i \binom{k}{i} \frac{1}{j!} E(X_{t+1}^{[j]}|X_t = i)$$

Felhasználva a $\sum_{i=0}^{k}(-1)^i i^m \binom{k}{i} = 0$ $(m < k)$, illetve $\sum_{i=0}^{k}(-1)^i i^k \binom{k}{i} = (-1)^k k!$ azonosságokat, és a feltétel átfogalmazását, ebből valóban adódik $a_{kj} = 0$ $(j < k)$, illetve $a_{kk} = \lambda_k$.

■

A tétel bizonyítására rátérve:

$$E(X_{t+1}^j|X_t = i) = E\left(\left(\sum_{k=1}^{i} Y_k\right)^j\right) = S_j(i)E\left(\prod_{k=1}^{j} Y_k\right) + i \text{ alacsonyabb fokú hatványai} =$$

$$= i^j E\left(\prod_{k=1}^{j} Y_k\right) + i \text{ alacsonyabb fokú hatványai,}$$

ahol csak a második egyenlőség igényel magyarázatot. Ha a j-dik hatvány minden tényezőjéből különböző Y_k-t szeretnék választani, azt $i^{[j]}$ féleképpen tehetem meg. Az olyan szorzatok száma pedig, ahol van olyan Y_k, ami 1-nél magasabb hatványon

szerepel, i-nek j-nél alacsonyabb fokú hatványa. Ezután a lemma alkalmazásával kapjuk az állítást. ■

A triviális $P(Y_l = 1) = 1$ esetben nyilván minden sajátérték 1. Más esetben ez nem fordulhat elő, sőt, pontosan meg tudjuk mondani a különböző sajátértékek számát is. Ehhez vezessük be a következő jelölést: jelölje k azt a természetes számot, amire igaz, hogy minden generációban $N-k+1$ egyedhez tartozó Y_l mindenképp 0. Azaz, minden lépésben $(2N-k+1)$ egyednek utód nélkül kell meghalnia, de $(2N-k+2)$-nek már nem szükséges.

5.3. Tétel. *Ha $P(Y_l = 1) = 1$ nem teljesül, akkor*

$$1 = \lambda_0 = \lambda_1 > \lambda_2 > \cdots > \lambda_k = \lambda_{k+1} = \cdots = \lambda_{2N} = 0.$$

Bizonyítás. Definíció szerint k darab Y_l közül az egyik értéke mindenképp 0 lesz, így $E(\prod_{l=1}^{j} Y_l) = 0$, így $\lambda_j = 0$, ha $j \geq k$.

$\lambda_0 = 1$ definíció szerint, míg $\lambda_1 = E(Y_1) = 1$ szimmetriaokokból teljesül. Ha $j < k$, akkor mivel $(k-1)$ darab Y_l pozitív valószínűséggel nem mind 0, ezért $\lambda_j = = E(\prod_{l=1}^{j} Y_l) > 0$. Így már csak a $\lambda_{j-1} > \lambda_j$ egyenlőtlenségeket kell belátni $1 < j < k$ esetén. Mivel $\sum_{l=1}^{2N} Y_l = 2N$, ezért

$$E(Y_1 Y_2 \cdots Y_j) = E(Y_1 Y_2 \cdots Y_{j-1}(2N - Y_1 - \cdots - Y_{j-1} - Y_{j+1} - \cdots Y_{2N})) =$$

$$= 2NE(Y_1 \cdots Y_{j-1}) - (j-1)E(Y_1 \cdots Y_{j-2} Y_{j-1}^2) - (2N-j)E(Y_1 \cdots Y_j)$$

a felcserélhetőség miatt. Átrendezés után

$$E(Y_1 \cdots Y_j) = \frac{2NE(Y_1 \cdots Y_{j-1}) - (j-1)E(Y_1 \cdots Y_{j-2} Y_{j-1}^2)}{2N-j+1}. \tag{21}$$

Az $Y_1 \cdots Y_{j-2} Y_{j-1}(Y_{j-1} - 1)$ valószínűségi változó nemnegatív, és mivel $P(Y_{j-1} = = 1) < 1$ illetve $j - 1 < k$, ezért pozitív valószínűséggel vesz fel pozitív értéket is, így

$$E(Y_1 \cdots Y_{j-2} Y_{j-1}(Y_{j-1} - 1)) > 0 \Leftrightarrow E(Y_1 \cdots Y_{j-2} Y_{j-1}^2) > E(Y_1 \cdots Y_{j-1}).$$

Ezt használva (21)-ben

$$\lambda_j = E(Y_1 \cdots Y_j) < E(Y_1 \cdots Y_{j-1}) = \lambda_{j-1}.$$

■

A legnagyobb nem 1 sajátérték tehát $E(Y_1Y_2)$, amit a következőképpen is megkaphatunk:

$$(2N)^2 = E\left(\left(\sum_{l=1}^{2N} Y_l\right)^2\right) = 2NE(Y_1^2) + 2N(2N-1)E(Y_1Y_2)$$

a felcserélhetőség miatt. Innen

$$E(Y_1Y_2) = 1 - \frac{E(Y_1^2) - 1}{2N-1} = 1 - \frac{\sigma^2}{2N-1},$$

ahol $\sigma^2 = D^2(Y_1)$.

Megjegyezzük, hogy a sajátértékekre ismert explicit formula mutáció, sőt, $M > 2$ allél esetén is, ezek azonban már jelentősen bonyolultabbak. Ekkor a λ_l sajátértékek a genetikai diverzitás csökkenésének sebességét jellemzik: ha a jelenlegi generációban l allél van, akkor annak a valószínűsége, hogy a következő generációban l-nél kevesebb allél van, aszimptotikusan $1 - \lambda_l$.

5.2. Speciális esetek

Az 5.1. Tételt alkalmazva kiszálmolhatjuk a korábbiakban tárgyalt két modell esetében az átmenetmátrix sajátértékeit. A Wright-Fisher modellben $(Y_1, Y_2, ..., Y_{2N})$ multinomiális eloszlású $2N$ renddel és egységesen $\frac{1}{2N}$ paraméterrel. Vezessük be a következő jelölést, ha $\mathbf{y} = (y_1, y_2, ..., y_j)$:

$$\binom{2N}{\mathbf{y}} = \frac{2N!}{y_1!\, y_2! \cdots y_j!\, (2N - y_1 - y_2 - \cdots y_j)!}.$$

Ekkor

$$\lambda_j = E(\prod_{l=1}^{j} Y_l) = \sum \cdots \sum y_1 \cdots y_j \binom{2N}{\mathbf{y}} \left(\frac{1}{2N}\right)^{\sum y_l} \left(1 - \frac{j}{2N}\right)^{2N - \sum y_l} =$$

$$= \frac{(2N)^{[j]}}{(2N)^j} \sum \cdots \sum \binom{2N-j}{\mathbf{y\text{-}1}} \left(\frac{1}{2N}\right)^{\sum (y_l - 1)} \left(1 - \frac{j}{2N}\right)^{2N - j - \sum (y_l - 1)} = \frac{(2N)^{[j]}}{(2N)^j},$$

hiszen az átalakított összegzés szintén egy multinomiális eloszlás valószínűségeinek az összegzése, azaz 1.

A Moran-modell esetében a rendezetlen $\{Y_1, ..., Y_{2N}\}$ halmaz két értéket vehet fel: {1,1, ...,1}-et $\frac{1}{2N}$ valószínűséggel, illetve {2,0,1,1, ...,1}-et $\frac{2N-1}{2N}$ valószínűséggel. Az első esetben $E(\prod_{l=1}^{j} Y_l | \{Y_1, ..., Y_{2N}\} = \{1,1, ...,1\}) = 1$, míg a másodikban $E(\prod_{l=1}^{j} Y_j | \{Y_1, ..., Y_{2N}\} = \{2,0,1,1, ...,1\}) = \frac{(2N-j)(2N+j-1)}{2N(2N-1)}$. Tehát

$$\lambda_j = \frac{1}{2N} + \frac{2N-1}{2N} \frac{(2N-j)(2N+j-1)}{2N(2N-1)} = 1 - \frac{j(j-1)}{(2N)^2} \qquad (j = 0,1, ...,2N).$$

6. Családfák vizsgálata

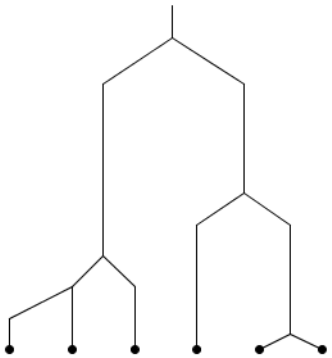
6.1. A retrospektív nézőpont

A korábbi fejezetekben azt vizsgáltuk, hogy különböző feltevések mellett a populáció bizonyos tulajdonságai az idő előrehaladtával hogyan változnak. A populáció jövőjének vizsgálata helyett ugyanakkor lehetséges a múltjával foglalkozni. A populációgenetika története során ez a megközelítés azért válhatott érdekessé, mert a technika fejlődésével egyre több adat állt rendelkezésre populációk genetikai állományáról, és felmerült a kérdés, hogy milyen folyamatok vezethettek a megfigyelt mintához.

Az ilyen irányú vizsgálatok egy fontos eszköze a családfák szerkezetéhez kapcsolódó *coalescent* folyamat. Vegyünk egy n elemű mintát az r-dik, G_r generációból, legyenek ezek az egyedek $\gamma_1, \gamma_2, \ldots \gamma_n$. A minta szülei G_{r-1}-ben egy legfeljebb n elemű halmazt alkotnak, de néhány szülő egybeeshet. Az egybeesés valószínűsége pozitív, így néhány generációt visszamenve, a mintának biztosan kevesebb mint n őse lesz. Ezt ismételve pedig előbb-utóbb egy olyan generációhoz is eljutunk, amikor az egész mintának egy közös őse van. Ezzel a minta családfáját kaptuk meg (1.ábra), az így kapott vonalakat családi vonalaknak, azok közös szülő miatti egybeesését összeolvadásnak is hívjuk. Ezt ekvivalenciarelációk segítségével is megfogalmazhatjuk: Ψ_s legyen az az ekvivalenciareláció $\{1,2,...n\}$-en, ami pontosan akkor tartalmazza az (i,j) párt, ha γ_i és γ_j őse G_{r-s}-ben megegyezik. Nyilván $\Psi_0 = \{(i,i), i = 1,2,...n\}$, és $\Psi_s \subseteq \Psi_{s+1}$. Minden Ψ_s ekvivalenciaosztálynak megfelel a G_{r-s} generáció egy tagja. Ha két ilyen egyed szülője megegyezik, akkor (és csak akkor) a két megfelelő ekvivalenciaosztály összeolvad Ψ_{s+1}-ben. A (Ψ_s) sorozat tehát Markov-láncot alkot az $\{1,2,...,n\}$ halmazon értelmezett ekvivalenciarelációkkal mint állapottérrel, és a $P_{\xi\eta} = P(\Psi_{s+1} = \eta | \Psi_s = \xi)$ átmeneti valószínűségekre $P_{\xi\eta} = 0$ hacsak nem $\xi \subseteq \eta$.

6.2. Alkalmazás a Wright-Fisher modellben

Tegyük most fel, hogy a populációnkban a Wright-Fisher modell szerinti kapcsolat van az egymást követő generációk között. Ekkor $P_{\xi\eta}$-t konkrétabban is ki tudjuk számolni. Jelöljük $\xi \prec \eta$-val, ha η-t ξ két osztályának egybeolvasztásával kapjuk.



1. ábra.

Ekkor

$$P_{\xi\eta} = \delta_{\xi\eta} + \frac{1}{2N} q_{\xi\eta} + O\left(\frac{1}{N^2}\right)$$

ahol

$$q_{\xi\eta} = \begin{cases} 1, & \text{ha } \xi \prec \eta \\ \frac{k(k-1)}{2}, & \text{ha } \xi = \eta \text{ és } |\xi| = k \\ 0, & \text{egyébként.} \end{cases}$$

Mátrixos alakban a P_{2N} átmenetmátrixra

$$P_{2N} = I + \frac{1}{2N} Q + O\left(\frac{1}{N^2}\right).$$

Ekkor megmutatható ([7]), hogy minden t-re $\lim_{N\to\infty} P_{2N}^{\lfloor 2Nt \rfloor} = \exp(tQ)$. Nem bizonyítjuk, de ekkor, ha R az a folytonos paraméterű folyamat, amire $R(t) = \Psi(\lfloor 2Nt \rfloor)$, akkor $N \to \infty$ esetén R eloszlásban tart egy folytonos paraméterű S Markov-folyamathoz a Q generátorral. Nagy N-re tehát lehet közelíteni a folyamatot S-sel, amire tehát $\xi \neq \eta$ esetén

$$\frac{P_{\xi\eta}(h)}{h} = \frac{P(S(t+h) = \eta | S(t) = \xi)}{h} \to q_{\xi\eta} \text{ ha } h \to 0.$$

A továbbiakban azt fogjuk figyelni, hogy mikor olvadnak össze családi vonalak, azaz, mikor csökken az ekvivalenciaosztályok száma. Legyen tehát $|\xi| = k$, ekkor

$$P(|S(t+h)| = k-1 | S(t) = \xi) = \sum_{\xi \prec \eta} P_{\xi\eta}(h) = \sum_{\xi \prec \eta} 1 \cdot h + o(h) = \frac{k(k-1)}{2} h + o(h),$$

tehát

$$P(|S(t+h)| = k-1 \mid \ |S(t)| = k) = \frac{k(k-1)}{2} h + o(h)$$

A coalescent tehát felfogható egy tiszta halálozási folyamatként, aminek a paraméterei $\frac{j(j-1)}{2}$ $(j = n, n-1, ...,2)$. Az ahhoz szükséges T_j idő, hogy j ősből $j-1$ legyen, exponenciális eloszlású $\frac{2}{j(j-1)}$ várható értékkel, és T_j-k függetlenek. Az egyetlen közös ős eléréséig eltelt T időre $T = T_2 + T_3 + \cdots + T_n$, így

$$E(T) = \sum_{i=2}^{n} E(T_i) = 2(1 - \frac{1}{n}). \tag{22}$$

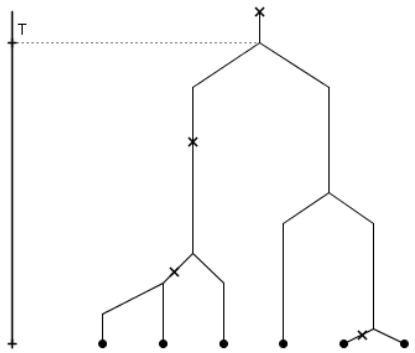
Mivel $E(T_2) = 1$, ezért az egyetlen közös ősig eltelt időnek átlagosan több mint a felében pontosan 2 közös őse van a mintának. Hasonlóan,

$$D^2(T) = \sum_{i=2}^{n} D^2(T_i) = 8 \sum_{i=1}^{n-1} \frac{1}{i^2} - 4\left(1 - \frac{1}{n}\right)\left(3 + \frac{1}{n}\right)$$

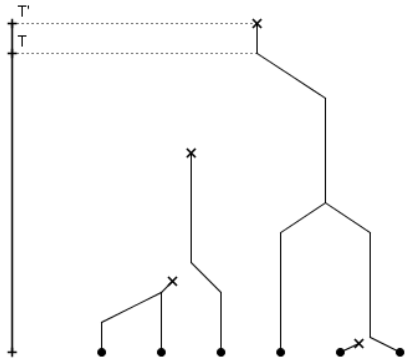
A szórás nagy részéért is a kevés őssel eltöltött idő szórása a felelős, például $D^2(T_6 + + T_7 + \cdots + T_n) \leq D^2(T_6 + T_7 + \cdots) < 0{,}011$.

A szülő-gyerek viszonyon kívül a folyamatban figyelembe vehetjük a mutációt is. Az eredeti Wright-Fisher modell-beli mutáció u valószínűséggel következett be egy génnél, 1 időegység alatt. A coalescent definiálása során az időt ugyanakkor átparamétereztük. Nem túl merész tehát most arra gondolni, hogy a mutáció valószínűsége 1 időegység alatt legyen $2Nu = \frac{\theta}{2}$. Mivel ez nem feltétlen értelmes, ezért ehelyett azt tesszük fel, hogy egy adott vonalon a $(t, t+\delta)$ intervallumban a mutáció valószínűsége legyen $\delta\frac{\theta}{2} + o(\delta)$, ha $\delta \to 0$.

A családfa vizsgálatakor tehát időről időre mutációkat figyelhetünk meg, amik szintén egy Poisson-folyamat szerint fordulnak elő, ezeket feljegyezhetjük a családfán



2. ábra.



3. ábra.

(2. ábra). Ha a t időpontban j közös ős van, akkor annak a valószínűsége, hogy a $(t, t+\delta)$ intervallumban mutáció vagy családi vonalak összeolvadása történik, $\frac{1}{2}j(j+$ $+\theta-1)\delta+o(\delta)$.

Először a mintában található legidősebb allél korát szeretnénk meghatározni. A továbbiakban, ha mutációt észlelünk egy családi vonalon, akkor azt a vonalat ne vizsgáljuk tovább (3. ábra). Így akár mutáció, akár összeolvadás történik, a vonalak száma 1-gyel csökken. Az előbbiek alapján, ha j vonal van, akkor a következő csökkenésig szükséges T_j' idő exponenciális eloszlású $\frac{1}{2}j(j+\theta-1)$ paraméterrel. A legidősebb allél

életkora $T' = \sum_{j=1}^{n} T'_j$, így

$$E(T') = \sum_{j=1}^{n} E(T'_j) = \sum_{j=1}^{n} \frac{2}{j(j+\theta-1)}. \tag{23}$$

Összehasonlítva (22)-t és (23)-t, az egyértelmű közös ősig $\theta < 2$ esetén várhatóan hamarabb eljutunk, mint a legidősebb allél létrejöttéig, így az ő típusa is jelen van a mintában. $\theta > 2$ esetén a fordított eset áll fenn, míg $\theta = 2$ esetén a két időpont csak egy $\frac{1}{n^2}$ nagyságrendű tagban különbözik.

Ha a Z_1 és Z_2 valószínűségi változók exponenciális eloszlásúak λ_1 illetve λ_2 paraméterekkel, akkor $P(Z_1 < Z_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$. Ha tehát bármilyen esemény (tehát mutáció vagy összeolvadás) történik, akkor annak a valószínűsége, hogy ez mutáció, $\frac{\theta}{j+\theta-1}$. A mintában lévő allélok számának várható értéke tehát

$$\sum_{j=1}^{n} \frac{\theta}{j+\theta-1},$$

ami megegyezik a 3.6. Következményben kapott eredménnyel. Ennél még többet is mondhatunk. Ismét K_n-nel jelölve a különböző allélok számát,

$$P(K_n = k) = \sum_{\substack{j_1,j_2,\ldots,j_{n-k} \\ j_l \neq j_m l \neq m, 1 \leq j_l \leq n}} \frac{\theta^k (j_1-1)(j_2-1)\cdots(j_{n-k}-1)}{\theta(\theta+1)\cdots(\theta+n-1)} =$$

$$= \frac{\theta^k}{S_n(\theta)} \sum_{\substack{j_1,j_2,\ldots,j_{n-k} \\ j_l \neq j_m l \neq m, 1 \leq j_l \leq n}} (j_1-1)(j_2-1)\cdots(j_{n-k}-1) = \frac{\theta^k |S_n^k|}{S_n(\theta)},$$

azaz a 3.5. Következményt is megkaptuk.

A családfa egészén megfigyelt mutációk S_n számának a statisztikai vizsgálatok során van fontos szerepe. Feltételezve ugyanis, hogy a megfigyelt szekvencia bármelyik helyén legfeljebb egyszer történhetett mutáció, S_n éppen azon helyek száma, ahol nem mindegyik mintaelem rendelkezik ugyanazzal a bázissal. Így a következők alapján a mintából torzítatlan becslést tudunk adni θ-ra. Legyen a családfa összhossza L_n, nyilván $L_n = \sum_{j=2}^{n} jT_j$. $L_n = l$ esetén S_n Poisson eloszlású, $\frac{\theta l}{2}$ paraméterrel. Így

tehát

$$E(S_n) = E(E(S_n|L_n)) = \frac{\theta}{2}E(\sum_{j=2}^{n} jT_j) = \frac{\theta}{2}\sum_{j=2}^{n} j\frac{2}{j(j-1)} = \theta\sum_{j=1}^{n-1}\frac{1}{j},$$

illetve

$$D^2(S_n) = E(D^2(S_n|L_n)) + D^2(E(S_n|L_n)) = \frac{\theta}{2}\sum_{j=2}^{n} jE(T_j) + \frac{\theta^2}{4}\sum_{j=2}^{n} j^2D^2(T_j) =$$

$$= \frac{\theta}{2}\sum_{j=2}^{n} j\frac{1}{j(j-1)} + \frac{\theta^2}{4}\sum_{j=2}^{n} j^2\frac{4}{j^2(j-1)^2} = \theta\left(\sum_{j=1}^{n-1}\frac{1}{j} + \frac{1}{j^2}\right).$$

Hivatkozások

[1] C. Cannings: The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. Journal of Applied Probability, Vol. 6, No. 2, 260-290(1974)

[2] W. J. Ewens: The sampling theory of selectively neutral alleles. Theoretical Population Biology, 3, 87-112(1972)

[3] W. J. Ewens, P. Joyce: Mathematical Population Genetics, Lecture Notes. http://www.cimat.mx/Eventos/xepe/Guanajuatowarrenpaul.pdf(2009)

[4] F. M. Hoppe: Pólya-like urns and the Ewens' sampling formula. Journal of Mathematical Biology, 20, 91-94(1984)

[5] S. Karlin, H. M. Taylor: A First Course in Stochastic Processes, Second Edition. Academic Press, New York(1975)

[6] S. Karlin, J. L. McGregor: Addendum to a paper of W. Ewens. Theoretical Population Biology, 3, 113-116(1972)

[7] J. F. C. Kingman: On the genealogy of large populations. Journal of Applied Probability, Vol. 19, 22-43(1982)