Stohastički modeli u genetici

Per, Valentina

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Valentina Per

Stohastički modeli u genetici

Diplomski rad

Osijek, 2020.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Valentina Per

Stohastički modeli u genetici

Diplomski rad

Mentor: doc. dr. sc. Danijel Grahovac

Osijek, 2020.

Sadržaj

1 Uvod

Znanost o nasljeđivanju nazivamo genetika, gdje nasljeđivanje predstavlja proces koji dovodi do sličnosti između roditelja i potomaka. Populacijska genetika dio je genetike koji se bavi proučavanjem genetske strukture populacije odnosno učestalosti alela i genotipova. Također, ona proučava kako se te učestalosti mijenjaju kroz vrijeme pod utjecajem glavnih evolucijskih sila. Ti se utjecaji i evolucijske promjene očituju tek na razini populacije pa je uzrok tih promjena potrebno potražiti na mikrorazini, odnosno na razini gena. Za opisivanje tih promjena izrazito su bitni stohastički modeli. Tako populacijska genetika kroz vrijeme upotrebljava razne stohastičke modele, odnosno koristi ih za modeliranje promjene frekvencije gena neke populacije od jedne do druge generacije.

U prvom poglavlju ćemo reći nešto više o osnovnim pojmovima genetike. Prisjetit ćemo se što su to geni, DNK te aminokiseline. Reći ćemo nešto više o kromosomima i nasljeđivanju. Drugo poglavlje će nas uvesti u početke matematičke teorije populacijske genetike. Prisjetit ćemo se pojmova kao što su evolucija, selekcija i specijacija. Vidjet ćemo i kada populacija dolazi u stanje Hardy-Weinbergove ravnoteže te što predstavlja genetički drift. U trećem poglavlju upoznat ćemo se sa najpoznatijim modelom reprodukcije u populacijskoj genetici, Wright - Fisherovim modelom. Navest ćemo njegove glavne pretpostavke te brojne rezultate vezane uz njega. Također u ovom poglavlju ćemo se dotaknuti i teorije koalescencije. Četvrto poglavlje upoznat će nas sa Wright - Fisherovim modelom uz prisutvo mutacija. Pokazat ćemo neke od osnovnih rezultata vezanih uz ovaj model. Model beskonačnog alela predstavit ćemo u petom poglavlju, gdje ćemo se također dotaknuti i Hoppeove urne te Ewensove formule uzimanja u uzorak. U posljednjem poglavlju obradit ćemo Moranov model, inačicu Wright- Fisherova modela u slučaju preklapajućih generacija.

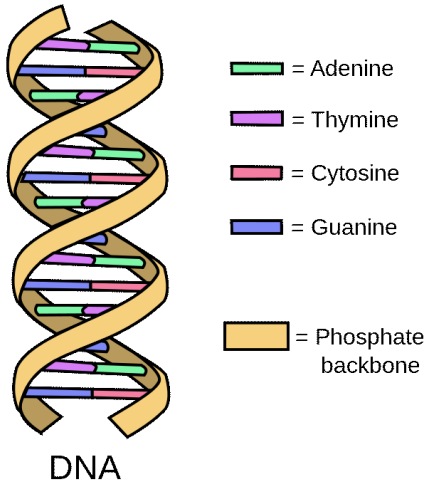
2 Genetika - osnovni pojmovi

U ovom ćemo se poglavlju upoznati sa osnovnim pojmovima vezanim za genetiku koji su nam potrebni za daljnje razumijevanje ovog rada. Pojmovi, definicije i primjeri navedeni u ovom poglavlju preuzeti su iz [8], [1], [5] i [9].

2.1 Geni i DNK

Geni predstavljaju kodiranu informaciju u obliku nasljedne jedinice koju potomci nasljeđuju od svojih roditelja. Oni su građeni od molekula deoksiribonukleinske kiseline, odnosno kraće DNK. Osnovna funkcija DNK je pohrana nasljednog materijala. Molekula DNK je dvolančana molekula koju čine dva polinukleotidna lanca omotana oko zamišljene osi u zavojnicu. Dva lanca molekule DNK međusobno su povezana vodikovim vezama.

Polinukleotidni lanac građen je od 4 različita nukleotida. Svaki nukleotid je građen od šećera deoksiriboze, fosfatne skupine i dušičnih baza. Dušične baze su A=adenin, G=gvanin, T=timin i C=citozin. Komplementarne parove [1] baza čine adenin i timin koji se međusobno spajaju dvjema vodikovim vezama odnosno citozin i gvanin koji se međusobno spajaju trima vodikovim vezama.



Slika 1: DNK dvostruka zavojnica. *Slika je preuzeta iz [1].*

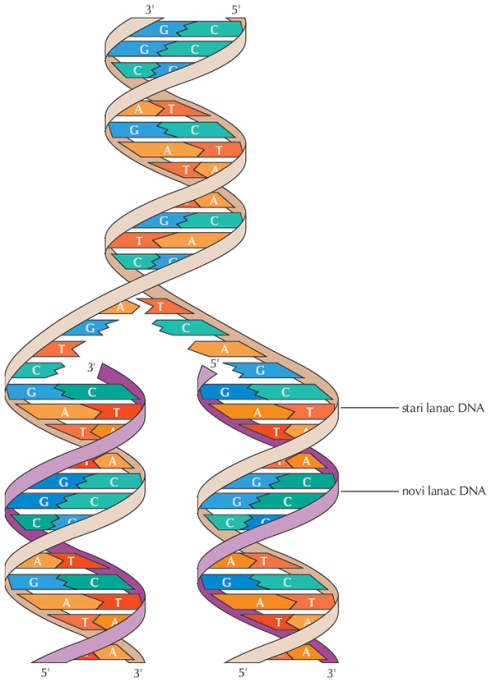
Primjer 2.1.1. Ukoliko nam je AGTC dio jedne spiralne niti onda je njemu ekvivalentan dio jednak TCAG.

Upravo ta specifičnost pri sparivanju baza je najvažnije svojstvo dvostruke zavojnice DNK. Slijed baza duž polinukleotidnog lanca nije ničime ograničen i on čini promjenjivi dio molekule. Redoslijed ovih baza nosi nasljednu informaciju i čini identitet određene osobe. Ljudski genom

[1] Lanci su suprotnog usmjerenja.

se sastoji od tri milijuna parova baza i to je najduži poznati polimerni lanac do danas. Kada je presavijena, molekula ljudske DNK duga je 6 metara.

DNK, upravo zbog svoje strukture dvostruke zavojnice i komplementarnosti lanaca omogućuje kopiranje i prijenos nasljednog materijala dijeljenjem kromosoma pri diobi stanice. Kada je DNK spremna za dijeljenje njezine dvije kromosomske niti se razdvajaju, odnosno razmataju. Kako u jezgri stanice ima slobodnih baza te baze se vežu na svoje parove na nitima DNK. Tako je informacija sačuvana, kopirana i prenešena na dva novonastala kromosoma.



Slika 2: Replikacija DNK lanca. *Slika je preuzeta iz [7].*

2.2 Aminokiseline

Aminokiseline su osnovni gradivni dijelovi staničnih struktura, proteina, hormona i enzima. Informacija o strukturi proteina je zapisana u kromosomu, odnosno u molekuli DNK. Svi proteini u svim organizmima, od bakterija do ljudi, građeni su od 20 aminokiselina. Nazive aminokiselina i njihove kratice možemo vidjeti u Tablici 1.

Genetska šifra sadržana u DNK mora biti na neki način zapisana linearnim redosljedom dušičnih baza duž polinukleotidnog lanca. Kao što smo već naveli lanci su komplementarni pa je dovoljno promatrati samo redoslijed u jednom lancu. Taj se redoslijed naziva genetički kod. Aminokiseline su kodirane trojkama susjednih nukleotida koji se nazivaju kodoni. Od 64 moguća kodona 61 kodon kodira aminokiseline, a preostala tri kodona UAA, UAG i UGA su stop kodoni koji označavaju završetak prepisivanja lanca. Na Slici 3. možemo vidjeti koja kombinacija u jednom tripletu predstavlja određenu aminokiselinu. Prvo slovo kodona je dano na lijevom rubu tablice, drugo na vrhu, a treće slovo na desnom rubu tablice.

Ala	Alanin	Leu	Leucin
Arg	Arginin	Lys	Lizin
Asn	Asparagin	Met	Metionin
Asp	Aspartat	Phe	Fenilalanin
Cys	Cistein	Pro	Prolin
Gly	Glicin	Ser	Serin
Glu	Glutamat	Thr	Treonin
Gln	Glutamin	Trp	Triptofan
His	Histidin	Tyr	Tirozin
Ile	Izoleucin	Val	valin

Tablica 1: Aminokiseline i njihove kratice

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Slika 3: Tablica genetske šifre. *Slika je preuzeta iz* [**??**].

Primjer 2.2.1. Prema prethodnoj tablici UAC predstavlja kod za Tirozin - Tyr.

2.3 Geni i nasljeđivanje

Geni se nalaze na strukturama koje se nazivaju kromosomi. Tijekom stanične diobe, svaka linearna molekula DNK se kondenzira u kromosom. Dakle, kromosomi su sastavljeni od gena, a geni su građeni od DNK. Kromosomi dolaze u parovima. Po dva kromosoma su jednake dužine,

centromere[2] su im na istom mjestu i nose gene za ista svojstva na istim položajima duž kromosoma. Takva dva kromosoma nazivamo homologni kromosomi, a čitava stanica (budući da ima dva seta kromosoma, odnosno po dva od svakog homolognog kromosoma) je diploidna, što označavamo sa $2n$. n predstavlja broj kromosoma u pojedinom setu (primjerice kod čovjeka $n = 23$, $2n = 46$). Stanica je haploidna ukoliko svaki homologni kromosom dolazi u samo jednom primjerku, odnosno stanica koja ima jedan set kromosoma. Niži organizmi poput bakterija su haploidni. Više detalja može se vidjeti u [9].

Kada dolazi do reprodukcije haploidnih organizama tada postoji jedan roditelj koji daje kopije svog genetskog materijala za svoje potomstvo. Kada se diploidni organizmi reproduciraju onda postoje dva roditelja pa potomak dobiva jedan kromosom od oca, a jedan od majke. U tom slučaju za svako svojstvo postoje dva gena ili dvije informacije. Dakle, svaki diploidni organizam ima po dva gena za isto svojstvo koji se nazivaju alel. Uvijek se nalaze na točno određenom položaju (lokusu) na homolognim kromosomima. Aleli su varijante gena. Možemo reći da su aleli linearni odsječci DNK molekule koji mogu imati identičan slijed nukleotida (onda ih označavamo istim slovom, npr. A). Također se njihovi slijedovi nukleotida mogu međusobno manje ili više razlikovati (onda ih označavamo različitim simbolima, npr. A i a). Ako organizam ima dva ista alela za neko svojstvo kažemo da je homozigot (npr. AA ili aa) za to svojstvo, a ako ima dva različita alela kažemo da je heterozigot (npr. Aa). Najčešći odnos alela je dominantno-recesivan. Jedan alel je dominantan, a drugi recesivan kada učinak dominantnog alela prikriva učinak recesivnog dok su aleli prisutni zajedno u heterozigotu. Prema konvenciji, dominantni alel se označava velikim slovom (npr. A), a recesivan malim (npr. a). No, napomenimo da recesivan alel ne mora uvijek biti i onaj koji je rjeđi u populaciji niti dominantan alel češći.

[2]Centromera ili kinetohora je pokretno središte kromosoma, mjesto na kojem kromosom za vrijeme stanične diobe prione uz diobeno vreteno.

3 Povijesna pozadina

U ovom ćemo se poglavlju upoznati sa početcima matematičke teorije populacijske genetike. Najprije ćemo nešto više reći o Hardy-Weinbergov zakonu kao polazištu u populacijskoj genetici. Također ćemo se pobliže upoznati sa pojmovima evolucije, selekcije, mutacije te genetičkog drifta. Pojmovi korišteni u ovom poglavlju preuzeti su iz [11], [16], [17] i [4].

3.1 Hardy-Weinbergov zakon

Pretpostavimo da je promatrana populacija vrlo velika te da nema fenotipskog[3] preferiranja pri izboru partnera za stvaranje potomstva. Također ukoliko se zanemari postojanje mutacija, migracija i genetičkog drifta te postojanje prirodne selekcije tada se populacija nalazi u stanju Hardy - Weinbergove ravnoteže. Pod tim bi uvjetima evolucija izostala, a da bi populacija evoluirala dovoljan je izostanak samo jednog od prethodno navedenih uvjeta. U stvarnosti je gotovo nemoguće postići Hardy - Weinbergovu ravnotežu. Razlog tome je što su aleli stalno podložni promjenama zbog nedostatka jednog ili više uvjeta. Ipak, Hardy-Weinbergov zakon ima smisla ukoliko neku populaciju promatramo u dovoljno kratkom vremenskom razdoblju, djelovanje evolucijskih sila je tako slabo da ga možemo zanemariti (evolucija se najčešće odvija tako sporo da je u kraćim vremenskim razmacima ne možemo primijetiti).

Hardy - Weinbergov zakon glasi: "Ako je populacija u ravnoteži, učestalosti alela, gena i genotipova ostaju nepromijenjene tijekom niza generacija."

Usredotočimo se sada na genetski lokus na kojem mogu postojati samo dvije varijante gena (dva alela) koji je odgovoran za točno određeno svojstvo. Označimo ih sa A i a. Tada se kao posljedica ovih pretpostavki na tom se lokusu može realizirati samo jedan od sljedećih genotipova AA, Aa i aa.
Oznake su sljedeće:

- AA - od oba roditelja je nasljeđen dominantni alel A;
- $Aa = aA$ - od jednog je roditelja nasljeđen dominantni alel, a od drugog recesivni alel a;
- aa - od oba roditelja je nasljeđen recesivni alel a.

Općenito, model za razdiobu genotipa u populaciji jednak je:

$$p_{AA} = P(\{AA\}),$$
$$p_{Aa} = P(\{Aa\}),$$
$$p_{aa} = P(\{aa\}),$$

pri čemu vrijedi: p_{AA}, p_{Aa}, $p_{aa} > 0$, $p_{AA} + p_{Aa} + p_{aa} = 1$. Pri tome vjerojatnosti ne moraju nužno biti jednake.

[3] Fenotip je skup svih genetski određenih svojstava jedinke.

Određivanje vrijednosti vjerojatnosti vezano je uz model razdiobe alela A i a u populaciji. Do tog ćemo modela doći ako genotip na promatranom lokusu prestanemo gledati kao cjelinu i shvatimo ga kao par alela od kojih je jedan nasljeđen od oca, a drugi od majke. Kako ne znamo koji je alel nasljeđen od kojeg roditelja, zaključujemo da su vjerojatnosti nasljeđivanja alela A, odnosno alela a, od istog roditelja jednake. Iz toga slijedi da vjerojatnosni model koji opisuje razdiobu alela u populaciji izgleda ovako:

$$p_A = P(\{A\}),$$
$$p_a = P(\{a\}),$$

pri čemu standardno vrijedi da su p_A, $p_a > 0$ i $p_A + p_a = 1$.

Uz poznate vjerojatnosti p_A i p_a iz modela razdiobe alela, zadovoljene uvjete Hardy - Weinbergove ravnoteže i nezavisnost alela nasljeđenih od pojedinog roditelja, lako slijedi:

1. $p_{AA} = \mathbb{P}(\{A\}, \{A\}) = \mathbb{P}(\{A\}) \cdot \mathbb{P}(\{A\}) = p_A^2$
2. $p_{Aa} = \mathbb{P}(\{\{A\}, \{a\}\} \cup \{\{a\}, \{A\}\}) = \mathbb{P}(\{A\}, \{a\}) + \mathbb{P}(\{a\}, \{A\}) = \mathbb{P}(\{A\}) \cdot \mathbb{P}(\{a\}) + \mathbb{P}(\{a\}) \cdot \mathbb{P}(\{A\}) = p_A \cdot p_a + p_a \cdot p_A = 2p_A p_a$
3. $\mathrm{p}_{aa} = \mathbb{P}(\{a\}, \{a\}) = \mathbb{P}(\{a\}) \cdot \mathbb{P}(\{a\}) = p_a^2$

Iz prikaza modela razdiobe genotipa, primjenom formule

$$p_{AA} + p_{Aa} + p_{aa} = 1$$

slijedi:

$$p_A^2 + 2p_A p_a + p_a^2 = 1,$$

pa prema formuli za kvadrat binoma imamo:

$$(p_A + p_a)^2 = 1.$$

Prethodna formula koristi se za izračunavanje udjela pojedinih alela u populaciji ako znamo da se ona nalazi u stanju Hardy - Weinbergove ravnoteže ili za provjeru uravnoteženosti populacije prema Hardy - Weinbergovom modelu ukoliko su nam poznati udjeli pojedinih alela. Za više detalja pogledati [11].

Primjer 3.1.1. Pretpostavimo da proučavamo boju krzna u populaciji miševa koja se nalazi u stanju Hardy -Weinbergove ravnoteže. Neka $p_A = 0.8$ predstavlja udio pojedinih alela koji određuju sivu (dominantno svojstvo određeno alelom A), odnosno $p_a = 0.2$ bijelu (recesivno svojstvo određeno alelom a) boju krzna.

Lako zaključujemo da su udjeli pojedinih genotipa u populaciji sljedeći:

$$p_{AA} = p_A^2 = 0.64,$$
$$p_{Aa} = 2p_A p_a = 0.32,$$
$$p_{aa} = p_a^2 = 0.04.$$

S obizrom da je siva boja krzna dominantno svojstvo određeno alelom A slijedi da će $0.64+0.32 = 0.96$, odnosno 96% miševa imati krzno sive, a njih 4% bijele boje.

Mendeljevo pravilo ukazuje na to da realizacija križanja $AA \times AA$ mora biti genotip AA. Pretpostavimo da je za svaku generaciju učestalost genotipa AA jednaka X, genotipa Aa jednaka $2Y$ te genotipa aa jednaka Z. Zaključujemo da je tada prilikom križanja $AA \times AA$ učestalost jednaka X^2, za $AA \times Aa$ jednaka $4XY$, itd. Obzirom da genotip AA možemo dobiti samo od križanja $AA \times AA$ (s ukupnom vjerojatnoću 1), $AA \times Aa$ (s ukupnom vjerojatnošću 1/2) i od $Aa \times Aa$ (s vjerojatnošću 1/4) tada je učestalost X' od AA u sljedećoj generaciji jednaka

$$X' = X^2 + \frac{1}{2}(4XY) + \frac{1}{4}(4Y^2) = (X+Y)^2.$$

Slično dobivamo i učestalost $2Y'$ od Aa te učestalost Z' od aa.

$$2Y' = \frac{1}{2}(4XY) + \frac{1}{2}(4Y^2) + 2XZ + \frac{1}{2}(4YZ) = 2(X+Y)(Y+Z),$$

$$Z' = \frac{1}{4}(4Y^2) + \frac{1}{4}(4YZ) + Z^2 = (Y+Z)^2.$$

Učestalost X", $2Y$" i Z" za sljedeće generacije dobivamo zamjenjujući X', $2Y'$ i Z' s X", $2Y$" i Z" i X, $2Y$ i Z s X', $2Y'$ i Z'.

Primjerice,

$$X'' = (X' + Y')^2 = (X+Y)^2 = X',$$

a slično se dobije da je i Y" $= Y'$, Z" $= Z'$.

Dakle, učestalost genotipa dobivena drugom generacijom ista je i u trećoj generaciji pa stoga i u svim narednim generacijama. Učestalosti koje imaju ovo svojstvo mogu se okarakterizirati kao oni koji zadovoljavaju odnos

$$(Y')^2 = X'Z'.$$

Ukoliko taj odnos vrijedi u prvoj generaciji, tako da $Y^2 = XZ$, tada ne samo da ne bi bilo promjena u učestalosti genotipova između druge i sljedeće generacije nego bi ove učestalosti bile iste kao i one u prvoj generaciji. Populacije za koje vrijedi $Y^2 = XZ$ kaže se da učestalosti genotipa zadovoljavaju oblik Hardy-Weinberga. Za više detalja pogledati [4].

Kao što smo već i prije naveli uvjeti koji populaciju dovode u stanje Hardy - Weinbergove ravnoteže gotovo nikada se ne poklope i idealizacija su stvarne situacije no korisni su određivanje procjene udjela određenih genotipova u populaciji.

3.2 Evolucija, selekcija i specijacija

Evolucija predstavlja proces razvoja i prilagodbe živih bića uvjetima okoliša u kojem obitavaju. Promjena frekvencije alela ili genotipova u populaciji odnosno promjena genetičke strukture populacije kroz veliki broj generacija je evolucija. U najširem smislu, to je proces u kojem nizom promjena ili razvojnih stupnjeva živi organizam ili skupina organizama stječe karakteristične morfološke i fiziološke značajke. Nakon što su se razradile i detaljnije pojasnile neke od temeljnih bioloških činjenica i pojava evolucija je postala središnja biološka znanost. Područje evolucije je osobito unaprijeđeno dostignućima populacijske genetike. Genetska ravnoteža se

narušava kada se mijenjaju uvjeti u populaciji ili okolišu, a njezinim narušavanjem započinje proces evoculije. Osnovne sile evolucije su mutacija, selekcija i genetički drift. To su procesi koji remete nasljednu ravnotežu. Za više detalja pogledati [16].

Mutabilnost je sposobnost promjene nasljednog materijala. Promjene u strukturi gena nazivamo genskim mutacijama. Različiti oblici mutacija proširuju genetsku raznolikost populacije, odnosno njezinih genskih zaliha. Mutacije su moguće kod svakog gena pa su izvor nasljedne varijacije. Mutante bivaju pojačane interakcijom s drugim genima, i taj posredni učinak još je značajniji za raznolikost genotipova i fenotipova, a njihov utjecaj na evoluciju veći od neposrednog učinka samoga mutiranog alela.

Prirodni odabir ili selekcija predstavlja prirodni izbor između nositelja nasljednih faktora. Darwin[4] selekciju objašnjava kao borbu za opstanak između pojedinih organizama od kojih preživljavaju najsposobniji. Dakle, Darwin selekciju primjenjuje na jedinke, a ne na populaciju. Danas je modificirano načelo selekcije temeljna orijentacija u rješavanju evolucijskih problema. Prema Hardy-Weinbergovu pravilu u standardnim uvjetima okoline svi geni populacijskih genskih zaliha dolaze do ravnoteže koja se stalno održava (idealna populacija). Selekcija je utjecaj bilo kojega faktora iz okoliša organizma. Primjerice selekcijski faktori mogu biti ekstremne temperature, sušna razdoblja, poplave, paraziti, različite bolesti, itd. Također se borba oko hrane, životnoga prostora i ostalih životnih uvjeta ubraja se u selekciju koja podređene skupine potiskuje ili vodi njihovu izumiranju. Tijekom godina selekcija omogućuje razvoj novih adaptacija u najrazličitijim sredinama na Zemlji. Više detalja može se vidjeti u [17].

Evolucija koja se odvija unutar populacije naziva se mikroevolucija. Ona sadrži manje sukcesivne promjene u genskoj zalihi određene populacije od jedne do druge generacije. U njoj djeluju osnovne sile evolucije, a procesi mikroevolucije vode stvaranju novih vrsta, tj. populacija. Nastanak novih vrsta, tj. odjeljivanje i razvoj novih populacija naziva se *specijacija*. U proučavanju mikroevolucijskih procesa populacijska genetika osigurava matematičku strukturu.

3.3 Genetički drift

Genetički drift odnosi se na promjene u učestalosti alela u genskoj zalihi koje su slučajnog karaktera. Kako je učestalost alela slučajnog karaktera te može rasti ili opadati zaključujemo da učestalost alela zapravo predstavlja relativnu frekvenciju alela. U većim populacijama je manja vjerojatnost da će doći do genetičkog drifta, dakle najvažniji faktor za genetički drift je veličina populacije. Kada je broj jedinki u populaciji mali, slučajnost može biti odlučujući faktor koji će odrediti koji će aleli u populaciji biti češći, a koji rjeđi. Genetički drift se odvija kad se slučajno određeni članovi populacije razmnožavaju i prenose gene u iduću generaciju.

Genetički drift uvjetuje ustaljivanje neutralnih ili neadaptivnih svojstava. Gubitak ili fiksacija događa se bez obzira na selekcijski pritisak. Osim selekcijskog istrjebljenja, postoji i mogućnost da se oni geni, inače samo usputni, koji se u velikoj genskoj zalihi nisu mogli istaknuti, odjednom istaknu u slučajem uvjetovanom genetskom driftu u malim populacijama i tako postanu

[4]https://hr.wikipedia.org/wiki/Charles_Darwin

valjani za neki genetski sastav, dok se oni geni koji su prije toga bili zastupljeni sada gube. Proces genetičkog drifta predstavlja važnu ulogu u molekularnoj evoluciji i ponašanju gena u populaciji s konačnim brojem jedinki. Pojam i tehnika nasumičnog genetičkog drifta imaju široko primjenu u medicini, matematici i drugim znanostima. Za više informacija o genetičkom driftu pogledati [15].

Nadalje ćemo se pozabaviti sa matematičkom formulacijom genetičkog drifta. Kako bi smo dobili osjećaj za genetički drift pretpostavit ćemo da slučajno odabiremo dvije vrste alela, A i a. Nadalje, neka p predstavlja vjerojatnost da bude odabran alel A, a neka $q = 1-p$ predstavlja vjerojatnost odabira alela a. Označimo veličinu populacije sa N te sa k označimo broj alela A. Prisjetimo se kako veličina populacije od N jedinki ima $2N$ kopija gena, tj. alela. Obzirom na ove informacije zanima nas kako možemo izračunati broj alela A, odnosno k? Neka je X diskretna slučajna varijabla kojom modeliramo broj alela u $2N$ kopija gena. Tada je vjerojatnost realizacije točno k alela A jednaka

$$P(X = k) = \binom{2N}{k} p^k q^{2N-k}.$$

Za slučajnu varijablu X kažemo da ima binomnu distribuciju s parametrima $2N$ i p, matematičko očekivanje $2Np$ te varijancu $2Npq$. Za više detalja posjetiti [14].

4 Wright - Fisherov model bez mutacija

U ovom poglavlju predstavit ćemo Wright - Fisherov model, jedan od daleko najpoznatijih modela reprodukcije u populacijskoj genetici. Pojmovi korišteni u tekstu ovog poglavlja temeljeni su na [2] i [10]. Prvo što moramo učiniti je pojasniti neke od pojmova i definicija koji su nužni za dobro razumijevanje ovog poglavlja.

Genetski lokus predstavlja lokaciju u genomu organizma.

Fitnes općenito predstavlja sposobnost nekoga pojedinca da preživi i reproducira se. Dakle, jedinke koje su sposobnije u preživljavanju i reprodukciji imaju veći fitnes.

Diploidne jedinke imaju dvije kopije svog genetskog materijala u svakoj stanici. U svakoj stanici postoje dva alela za određeno svojstvo pa stoga u slučaju kada imamo N stanica zapravo imamo N parova alela, tj. $2N$ alela.

Proces slučajnog razmnožavanja predstavlja proces u kojem svaka jedinka neke populacije ima jednaku vjerojatnost razmnožavanja sa bilo kojom drugom jedinkom populacije suprotnoga spola (engl. random mating).

Genski bazen predstavlja sve gene, ukupnu genetičku raznolikost populacije jedne vrste (eng. gene pool).

4.1 Slučajni procesi i osnovne definicije

Sljedeće definicije preuzete su iz [12] i [13] .

Definicija 4.1.1. Familija σ-podalgebri $\mathbb{F} = (\mathcal{F}_n, n \in \mathbb{N}_0)$ σ-algebre $\mathcal{F}$ sa svojstvom

$$\mathcal{F}_n \subset \mathcal{F}_{n+1}, \quad \forall n \in \mathbb{N}_0$$

naziva se filtracija na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Vjerojatnosni prostor $(\Omega, \mathcal{F}, P)$ opskrbljen filtracijom $\mathbb{F}$ nazivamo filtriranim vjerojatnosnim prostorom te uvodimo oznaku $(\Omega, \mathcal{F}, P, \mathbb{F})$.

Definicija 4.1.2. Slučajni proces $X = (X_n, n \in \mathbb{N}_0)$ na filtriranom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P, \mathbb{F})$ je martingal u diskretnom vremenu ako zadovoljava sljedeće zahtjeve:

1. $E[|X_n|] < \infty$ za sve $n \in \mathbb{N}_0$
2. X je $\mathbb{F}$-adaptiran slučajni proces
3. $E[X_{n+1}|\mathcal{F}_n] = X_n$, za sve $n \in \mathbb{N}_0$.

Definicija 4.1.3. Neka je S diskretan skup stanja. Slučajni proces $X = (X_n, n \in \mathbb{N}_0)$ na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ s vrijednostima u skupu S je Markovljev lanac (ML) ako jednakost

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0) = P(X_{n+1} = j | X_n = i) \qquad (1)$$

vrijedi za svaki $n \in \mathbb{N}_0$ i za sve $i_0, \dots i_{n-1}$, $i, j \in S$ za koje su uvjetne vjerojatnosti u (1) dobro definirane.

Svojstvo (1) zove se Markovljevo svojstvo (MS), a možemo ga interpretirati na sljedeći način: *vjerojatnosno ponašanje procesa X u neposrednoj budućnosti, uvjetno na sadašnjost i prošlost, jednako je njegovom ponašanju u neposrednoj budućnosti uvjetno samo na sadašnjost.*

Markovljevi lanci u diskretnom vremenu su klasa procesa za koje vjerojatnost prijelaza iz jednog u drugo stanje zadajemo funkcijom prijelaznih vjerojatnosti - za stanja $i, j \in S$ i trenutke $s, t \in \mathbb{N}_0$ t.d. je $s < t$ **funkciju prijelaznih vjerojatnosti** definiramo pravilom

$$p(i, s; t, j) = P(X_t = j | X_s = i). \tag{2}$$

Za konkretna stanja $i, j \in S$ i trenutke $s, t \in \mathbb{N}_0$ t.d. je $s < t$ vrijednost funkcije $p(i, s; t, j)$ interpretiramo na sljedeći način: ako se u trenutku s ML nalazio u stanju i, $p(i, s; t, j)$ je vjerojatnost da se u trenutku t nade u stanju j, tj. prijede u stanje j.
Funkcija prijelaznih vjerojatnosti u jednom koraku (funkcija 1-koračnih prijelaznih vjerojatnosti) za stanja $i, j \in S$ i $n \in \mathbb{N}_0$ definirana je pravilom

$$p(i, n; n+1, j) = P(X_{n+1} = j | X_n = i). \tag{3}$$

Definicija 4.1.4. Za skup $C \subset S$ kažemo da je zatvoren podskup skupa stanja S ako je za svaki $i \in C$

$$P(T_{S \setminus C} = \infty | X_0 = i) = 1,$$

gdje je $T_{S \setminus C}$ prvo vrijeme izlaska Markovljevog lanca iz skupa $C \subset S$.

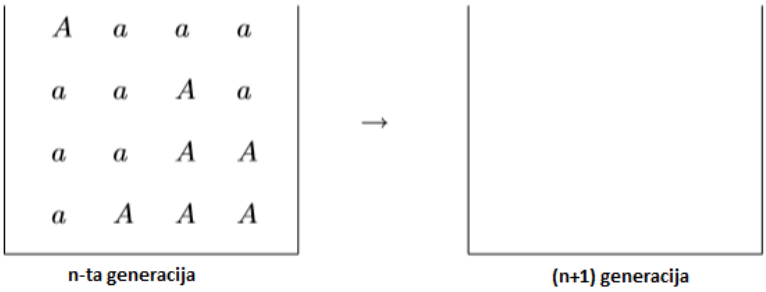
Definicija 4.1.5. Za stanje $j \in S$ kažemo da je apsorbirajuće ako je j zatvoren podskup skupa S.

Nakon što smo se upoznali sa nekim od osnovnih definicija pogledajmo Wright - Fisherov model bez prisutnih mutacija u populaciji. Wright - Fisherov model počiva na nekim pretpostavkama. To su:

- *Nepreklapajuća generacija.*
- *Haploidne jedinke.*
- *Veličina populacije je konstanta.*
- *Sve jedinke imaju isti fitnes.*
- *Populacija nema geografsku ili socijalnu strukturu.*
- *U populaciji nema rekombinacije gena.*

Za više detalja pogledati [6].

Stoga ćemo započeti sa promatranjem genetskog lokusa s dva alela A i a koji imaju jednak fitnes u diploidnoj populaciji konstantne veličine N s generacijama koje se ne preklapaju i koje prolaze kroz proces slučajnog razmnožavanja.



Slika 4: Wright - Fisherov model. *Slika je preuzeta iz [2].*

Sa Slike 4 vidimo da stanje populacije u početnoj generaciji n možemo promatramo kao genetski bazen koja se sastoji od $2N$ alela. Neka i predstavlja broj onih alela koje smo označili s A pa je stoga $2N - i$ broj onih alela koje smo označili s a. Sljedeću, odnosno $(n+1)$ generaciju dobivamo nasumičnim odabirom $2N$ alela. Nakon svakog odabira izvučeni alel se vraća u roditeljsku populaciju. Može se dogoditi da neki od alela nema potomaka u $(n+1)$ generaciji pa je stoga njegovo nasljedstvo izumrlo.

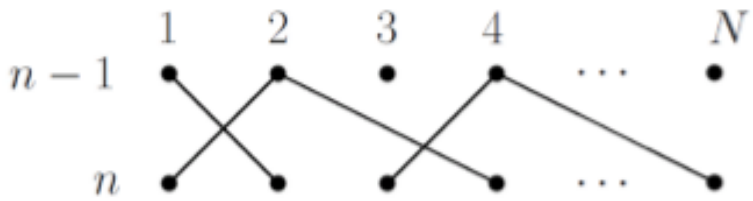
S obzirom na definiciju binomne distribucije vidimo da je vjerojatnost da u trenutku $(n+1)$ imamo j alela A kada u trenutku n imamo i alela A jednaka

$$p(i,j) = \binom{2N}{j} p_i^j \left(1 - p_i\right)^{2N-j}, \quad i,j = 0,1,\ldots,2N. \tag{4}$$

Izrazom $p_i = \frac{i}{2N}$ dana je vjerojatnost da u trenutku $(n+1)$ u jednom pokušaju slučajno bude obabran alel tipa A s obzirom da postoji i alela tipa A u trenutku n. Također, prisjetimo se da je

$$\binom{2N}{j} = \frac{(2N)!}{j!(2N-j)!}$$

binomni koeficijent koji predstavlja broj svih kombinacija j-tog razreda u $2N$-članom skupu.



Slika 5: Primjer slučajne reprodukcije u Wright - Fisherovom modelu. *Slika je preuzeta iz [1].*

Pretpostavimo da u svakom trenutku svaka jedinka slučajnim odabirom odabire drugu jedinku (moguće i sebe samog). Neka je X_n, $n \in \mathbb{N}$ slučajna varijabla koja modelira broj alela A u n-toj generaciji. Lako uočavamo da je proces $\{X_n, n \in \mathbb{N}_0\}$ Markovljev lanac zbog toga što slučajna varijabla X_{n+1} koja modelira broj alela u $(n+1)$ generaciji ovisi samo o broju alela u

n-toj populaciji, odnosno s obzirom na sadašnje stanje, prošlost je irelevantna za predviđanje budućnosti. Stoga zaključujemo da je X_n Markovljev lanac sa skupom stanja S= {0,1,. . . , 2N} i funkcijom prijelaznih vjerojatnosti u jednom koraku

$$p(i,j) = P(X_{n+1} = j | X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \tag{5}$$

Ovaj izraz zapravo predstavlja vjerojatnost da će u $(n+1)$-oj generaciji biti j alela A ako ih je u n-toj generaciji bilo i.

Promotrimo sada ponašanje Wright - Fisherovog modela *tijekom dužeg vremenskog perioda*. Moguća su dva scenarija:

1. Broj alela A u generaciji n jednak je 0, što ukazuje na gubitak alela A.
2. Broj alela A u generaciji n jednak je $2N$, što ukazuje na gubitak alela a.

Kada se neki alel izgubi iz populacije on se u nju više nikad ne vraća. Stoga imajmo na umu da kada Markovljev lanac uđe u jedno od stanja on ih više nkada ne može napustiti. Vidimo da su 0 i $2N$ apsorbirajuća stanja ovog Markovljevog lanca te vrijedi $p(0,0) = p(2N,2N) = 1$. Definirajmo vrijeme apsorbcije na sljedeći način

$$\tau = min\{n : X_n = 0 \quad \text{ili} \quad X_n = 2N\}, \tag{6}$$

odnosno to je trenutak u kojem populacija sadržava sve alele tipa a ili sve alele tipa A.

Zanima nas vrijeme apsorbcije u svim A stanjima. Odgovor na to će nam dati sljedeći teorem koji je zajedno sa dokazom preuzet iz [2]. Napomenimo da P_i koristimo za označavanje vjerojatnosti procesa X_n polazeći od $X_0 = i$, a E_i za označavanje očekivane vrijednosti u odnosu na P_i.

Teorem 4.1. U Wright - Fisherovom modelu vjerojatnost apsorbcije u svim stanjima A jednaka je

$$P_i(X_\tau = 2N) = \frac{i}{2N}.$$

Dokaz. Kako je broj jedinki konačan i uvijek je moguće ili izvući sve alele A ili sve alele a, apsorbcija će se uvijek dogoditi. Neka X_n označava broj alela A u trenutku n. S obzirom na izraz (4) vidimo da je očekivanje broja alela A u $(n+1)$ generaciji ako je u generaciji n bilo i alela A jednako $2Np_i$. Tada imamo da je

$$E(X_{n+1}|X_n = i) = 2Np_i = 2N \cdot (\frac{i}{2N}) = i.$$

Možemo zaključiti da je ovo poznata vrijednost varijable X_n, odnosno

$$E(X_{n+1}|X_n = i) = X_n. \tag{7}$$

Iskoristimo li svojstvo uvjetnog očekivanja na izraz (7) dolazimo do zaključka da je

$$EX_{n+1} = EX_n.$$

Riječima, matematičko očekivanje slučajne varijable X_n konstantno je u vremenu. Kako je skup stanja S konačan imamo da je $P(\tau < \infty) = 1$ i $p(0,0) = p(2N, 2N) = 1$ slijedi da je

$$\lim_{n\to\infty} X_n = X_\tau.$$

Dalje, iterativnim postupkom na izraz (7) dobivamo

$$i = E(X_n|X_0 = i) = E_i(X_n|X_0 = i)P(\tau \leq n) + E_i(X_n|X_0 = i)P(\tau > n).$$

Uočimo da za $\tau \leq n$ imamo da je $X_n = X_\tau$ jer je apsorbirajuće stanje dostignuto u τ. Tada slijedi da je

$$i = E(X_n|X_0 = i) = E_i(X_\tau|X_0 = i)P(\tau \leq n) + E_i(X_n|X_0 = i)P(\tau > n). \tag{8}$$

Ukoliko $n \to \infty$ i iskoristimo li činjenicu da je $X_n \leq 2N$ dolazimo do zaključka da prvi dio izraza konvergira u $E_i X_\tau$, a drugi dio izraza konvergira u 0. Slijedi da je

$$E(X_\tau|X_0 = i) = E_i X_\tau = i. \tag{9}$$

Izjednačavanjem izraza (8) i (9) imamo

$$\begin{aligned} i &= 2N \cdot P_i\{X_\tau = 2N\} + 0 \cdot P_i\{X_\tau = 0\} \\ &= 2N \cdot P_i\{X_\tau = 2N\}, \end{aligned}$$

te sređivanjem izraza dobivamo

$$P_i(X_\tau = 2N) = \frac{i}{2N}.$$

■

Zanima nas koliko je vremena potrebno da se dogodi apsorbcija? U tom slučaju je potrebno ispitati *heterozigotnost*.

4.2 Heterozigotnost

Heterozigotnost predstavlja prisustvo različitih alela na istom genskom lokusu. Dakle na heterozigotnost možemo gledati kao na vjerojatnost da su dva slučajno odabrana alela, bez vraćanja, u trenutku n različita. Pojednostavljeno, vjerojatnost da se izaberu aleli A i a ili a i A. Neka je H_n^0 slučajna varijabla kojom modeliramo heterozigotnost. H_n^0 modeliramo kao

$$H_n^0 = \frac{2X_n(2N - X_n)}{2N(2N-1)}.$$

Pojasnimo ovaj izraz. Prisjetimo se ponovno genetskog bazena koji se sastoji od $2N$ alela. S obzirom da sa X_n označavamo broj alela A u trenutku n, onda nam $(2N - X_n)$ predstavlja slučajnu varijabla koja označava broj alela a u trenutku n. Tada prema principu produkta imamo da je $2X_n(2N - X_n)$ slučajna varijabla kojom modeliramo broj načina na koji iz genetskog bazena možemo izabrati dva različita alela. Isto tako je $2N(2N-1)$ broj načina na koji iz genetskog bazena, koji se sastoji od $2N$ alela, možemo odabrati bilo koja dva alela.

Teorem koji slijedi i njegov dokaz preuzeti su iz [2].

Teorem 4.2. Označimo sa

$$h(n) = EH_n^0$$

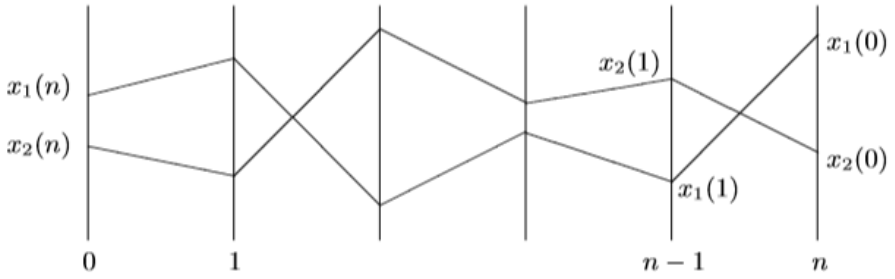
matematičko očekivanje heterozigotnosti u trenutku n. Tada u Wright - Fisherovom modelu vrijedi

$$h(n) = (1 - \frac{1}{2N})^n \cdot h(0) \tag{10}$$

Dokaz. Promotrimo $2N$ kopija lokusa $1, 2, \ldots, 2N$ i neka svaka od tih kopija bude promatrana kao jedinka. Pretpostavimo da u trenutku n odaberemo dvije jedinke. Označimo ih sa $x_1(0)$ i $x_2(0)$. Svaka od tih jedinki $x_i(0)$ je potomak neke jedinke $x_i(1)$ u trenutku $(n-1)$. Jedinka $x_i(1)$ je potomak neke jedinke $x_i(2)$ u trenutku $(n-2)$, itd. Vidimo da $x_i(m)$, za $0 \leq m \leq n$ opisuje podrijetlo $x_i(0)$, odnosno sve njegove pretke unazad tijekom vremena.
Ukoliko je $x_1(m) = x_2(m)$, onda je i $x_1(l) = x_2(l)$, za $m < l \leq n$.
Ako je $x_1(m) \neq x_2(m)$, onda su izbori dva roditelja napravljeni nezavisno, tako da je $x_1(m+1) \neq x_2(m+1)$ sa vjerojatnošću

$$1 - \frac{1}{2N}. \tag{11}$$

To slijedi iz činjenice da je vjerojatnost poududaranja dvije jedinke na istoj poziciji $\frac{1}{2N}$ pa korištenjem vjerojatnosti suprotnog događaja dolazimo do izraza (11). Da bi $x_1(n) \neq x_2(n)$ različiti roditelji bi trebali biti odabrani u svakom trenutku $1 \leq m \leq n$ sa vjerojatnošću događaja $(1 - \frac{1}{2N})^n$.
Kada se dvije linije ne podudaraju niti na jednoj od pozicija $1, 2, \ldots, 2N$, tada su $x_1(n)$ i $x_2(n)$ dvije slučajno odabrane jedinke iz populacije u trenutku 0. Stoga je vjerojatnost da su različite jednaka $H_0 = h(0)$. ■



Slika 6: Parovi evolucijskih linija. *Slika je preuzeta iz [2].*

Kada se jedinke nalaze na različitim mjestima, odnosno kada se ne podudaraju na poziciji i, $i \in \{1, 2, \ldots, 2N\}$, one se kreću nezavisno. Ukoliko se jedinke nađu na istom mjestu one se spoje i postaju jedna jedinka.

4.3 Teorija koalescencije

Teorija koalescencije otkrivena je od strane nekoliko istraživača u 1980-tim, ali definitivna formalizacija pripisuje se Sir John Kingmanu. Zbog toga se često naziva i Kingmanova koalescencija,

a u literaturi se o njoj govori kao aproksimaciji Wright - Fisherovog modela za velike populacije. Korist teorije koalescencije u kartiranju[5] gena za bolest sve se više primjenjuje. Danas brojni istraživači koji aktivno razvijaju algoritme za analizu genetičkih podataka čovjeka, koriste teoriju koalescencije.
Teorija koalescencije opisuje obiteljsko stablo velike haploidne populacije čije porijeklo se prati unazad. Prema toj teoriji doći će se do točke u vremenu u kojoj se nalazi Most recent common ancestor(MRCA) ili najbliži zajednički predak svih jedinki populacije. Mi na populaciju možemo gledati kao na genski bazen. Pretpostavimo da pratimo alele unazad kroz vrijeme. Kako se krećemo unazad u vremenu, broj predaka se počinje smanjivati sve dok ne ostanemo s jedinkom iz kojeg potječe čitava populacija. Ta se jedinka nalazi u korijenu koalescentnog stabla i zove se MRCA. MRCA ima svojstvo takav da je zajednički predak cijele populacije. Vraćajući se unatrag u vremenu, filogenetičke linije[6] se sjedinjuju (koalesciraju) kada god dvije ili više jedinki imaju istog pretka. One nikad ne završavaju. Dakle, prema teoriji koalescencije svi su aleli i geni neke populacije naslijeđeni od samo jednog pretka. Za više detalja pogledati [10].

Vrijeme koalescencije.
Sljedeći pojmovi preuzeti su iz [2] i [10].
Analiza temeljena na teoriji koalescencije traži predikciju količine vremena koje je proteklo između uvođenja mutacije i distribucije određenog gena ili alela u populaciji. Taj vremenski period jednak je vremenu u kojem je živio najbliži zajednički predak.

Vjerojatnost da dvije linije koalesciraju u prvoj neposrednoj prethodnoj generaciji jednaka je vjerojatnosti da oni imaju zajedničkog roditelja. U diploidnoj populaciji konstantne veličine s $2N$ kopija svakog lokusa, ima $2N$ potencijalnih roditelja u prethodnoj generaciji, dakle, vjerojatnost da dva alela imaju zajedničkog roditelja je $\frac{1}{2N}$ i shodno tomu, vjerojatnost da oni NE koalesciraju je $(1-\frac{1}{2N})$.

U svakoj uzastopnoj prethodnoj generaciji, vjerojatnost koalescencije je geometrijski distribuirana, znači, to je vjerojatnost NE koalescencije u $(n-1)$ prethodnih generacija multiplicirano s vjerojatnošću koalescencije u generaciji koja nas zanima:

$$P_c(n) = \left(1 - \frac{1}{2N}\right)^{n-1} \cdot \left(\frac{1}{2N}\right).$$

Za dovoljno velike vrijednost n, ova distribucija može se dobro aproksimirati eksponencijalnom distribucijom:

$$f_c(n) = (\frac{1}{2N}) \cdot e^{\frac{-n}{2N}}$$

Eksponencijalna distribucija ima očekivanu vrijednost i standardnu devijaciju jednaku $2N$. Iako je očekivano vrijeme koalescencije $2N$, stvarna vremena koalescencije imaju širok raspon vari-

[5]Kartiranje gena je proces kojim identificiramo i određujemo apsolutni odnosno relativni položaj i redoslijed genskog lokusa na kromosomu.

[6]Poznate i kao evolucijske linije predstavljaju slijed i međusobne odnose potomaka zajedničkoga pretka koji nastaju jedan iz drugoga grananjem.

jacije.

Znamo da za dovoljno mali x vrijedi da je $(1-x) \approx e^{-x}$. Stoga, kada je N velik mi izraz (10) možemo zapisati kao

$$h(n) \approx e^{-\frac{n}{2N}} \cdot h(0).$$

Ako promatramo k jedinki, vjerojatnost da će neke dvije od k jedinki imati istog roditelja iz prethodne generacije, tj. vjerojatnost sudara čestica je približno jednaka

$$\approx \frac{k(k-1)}{2} \cdot \frac{1}{2N}.$$

U prethodnom zapisu, izraz $\frac{k(k-1)}{2}$ predstavlja broj načina odabira dvije od k jedinki, a izraz $\frac{1}{2N}$ vjerojatnost da će dvije od k jedinki odabrati istog roditelja. Napomenimo da ovdje zanemarujemo vjerojatnost da dva različita para izaberu iste roditelje u jednom koraku ili da će tri jedinke izabrati istog roditelja.

Označimo sa T_j vrijeme u kojem se pojavljuje j genskih veza. Nadalje neka t_j označava vrijeme tijekom kojeg postoji točno j genskih veza. Mi možemo proces spajanja čestica i onog što se događa unazad u vremenu vizualizirati slikom. Radi jednostavnosti ne prikazujemo kako se linije kreću prije sudara, već prikazujemo samo vrijeme kada će se sudar, tj. sjedinjene dogoditi.

Ako je t_5 vrijeme tijekom kojeg postoji točno 5 genskih veza, nakon što dvije jedinke koje imaju zajedničkog pretka koalesciraju, t_4 je vrijeme tijekom kojeg postoji točno 4 genskih veza. Također, sa Slike 7 vidimo da je T_1 MRCA, tj. najbliži zajednički predak svih jedinki populacije. Sljedeći teorem i njegov dokaz preuzeti su iz [2].

Teorem 4.3.1. Kada mjerimo u jedinicama $2N$ generacija, vrijeme tijekom kojeg postoji k genskih veza, t_k, ima približno eksponencijalnu distribuciju sa očekivanjem

$$\frac{2}{k(k-1)}.$$

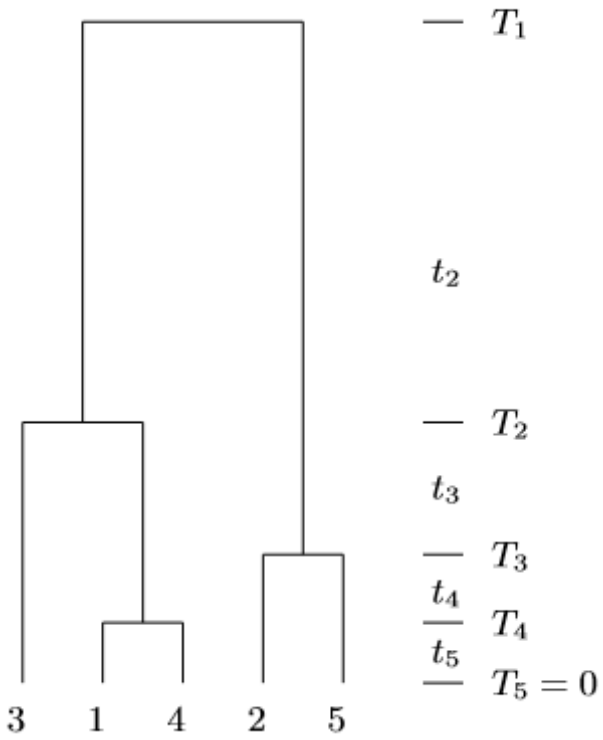
Dokaz. Kao što smo već naveli, vjerojatnost da će k jedinki imati istog roditelja, tj. vjerojatnost da će doći do sudara $\approx \frac{k(k-1)}{2} \cdot \frac{1}{2N}$. Tada primjenom vjerojatnosti suprotnog događaja, vjerojatnost da neće doći do sudara čestica u prvih n generacija približno je jednaka

$$\approx \left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}.\right)^n \approx exp\left(-\frac{k(k-1)}{2} \cdot \frac{n}{2N}\right).$$

Znamo da je eksponencijalna distribucija s parametrom λ definirana funkcijom distribucije

$$P(T \leq t) = \left(1 - e^{-\lambda t}\right) \cdot I_{(0,\infty)}(t)$$

i ima očekivanje $\frac{1}{\lambda}$. Ukoliko mi vrijeme izrazimo u terminima $2N$ generacija, onda je $t = \frac{n}{2N}$. Tada ako veličina populacije $N \to \infty$, onda vrijeme do prvog sudara ima pribiližno eksponencijalnu distribuciju sa očekivanjem $\frac{2}{k(k-1)}$. Koristeći se terminologijom teorije Markovljevih lanaca k čestica koalescira s $k-1$ čestica po stopi $\frac{2}{k(k-1)}$. Budući da se ovo obrazloženje primjenjuje u bilo kojem trenutku u kojem postoji k linija, slijedi željeni rezultat. ■



Slika 7: Realizacija procesa sjedinjena čestica za uzorak veličine 5. *Slika je preuzeta iz [2].*

Za uzorak veličine n ukupno vrijeme potrebno da uzorak populacije dođe do jedinke koja predstavlja zajednički predak cijelog uzorka jednako je $T_1 = t_n + \cdots + t_2$. Tada je očekivanje jednako

$$ET_1 = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2 \cdot \left(1 - \frac{1}{n} \right). \tag{12}$$

Izraz (12) konvergira ka 2 kada $n \to \infty$, ali vrijeme t_2 u kojem postoje samo dvije genske veze ima $Et_2 = 2 \cdot (1 - \frac{1}{2}) = 1$, tako da je očekivano vrijeme utrošeno čekajući posljednju koalescenciju uvijek čini barem polovicu ukupnog vremena koalescencije.

Oblik genealoškog stabla.
Mi genetsko stanje populacije u bilo koje vrijeme možemo prikazati kao particiju, $A_1, \ldots, A_m$ od $\{1, 2, \ldots, n\}$ te tada vrijedi $\cup_{i=1}^{m} A_i = \{1, 2, \ldots, m\}$. Ako je $i \neq j$ onda su skupovi A_i i A_j disjunktni. Riječima rečeno, svaki se A_i sastoji od jednog podskupa linija koje su se sjedinile. Za bolje razumijevanje poslužit ćemo se sa Slikom 7. U ovom slučaju, gledano unazad u vremenu, particije su

T_1	{1,2,3,4,5}				
T_2	{1,3,4}	{2,5}			
T_3	{1,4}	{2,5}	{3}		
T_4	{1,4}	{2}	{3}	{5}	
vrijeme 0	{1}	{2}	{3}	{4}	{5}

U početku ili kako smo označili vremenu 0, particija se sastoji od 5 jednočlanih skupova zbog toga što još nije došlo do spajanja, tj. koalescencije. Nakon što se 1 i 4 spoje u vremenu T_4 oni se pojavljuju u istom setu. Zatim se 2 i 5 sjedinjuju u vremenu T_3, itd. Konačno, u vremenu T_1 završimo sa svim jedinkama u jednom, istom skupu.

Za više detalja pogledati [2].

Neka je ε_n skup particija od $\{1, 2, \ldots, n\}$. Ako je $\xi \in \varepsilon_n$, neka je $|\xi|$ broj skupova koji čine ξ, tj. broj veza koje ostaju spojene. Ako, npr., $\xi = \{\{1\}\{2,3\},\{4,5\}\}$, tada je $|\xi| = 3$. Neka je ξ_i^n, $i = n, n-1, \ldots, 1$ particija od $\{1, 2, \ldots, n\}$ u vremenu T_i, prvo vrijeme u kojem postoji i veza.

Teorem 4.3.2. Ako je ξ particija od $\{1, 2, \ldots, n\}$, gdje je $|\xi| = i$, onda vrijedi

$$P(\xi_i^n = \xi) = c_{n,i} \cdot \omega(\xi). \tag{13}$$

U izrazu (13) za težine $\omega(\xi)$ vrijedi da je

$$\omega(\xi) = \lambda_1! \cdots \lambda_i!,$$

gdje su $\lambda_1, \ldots, \lambda_i$ veličine i skupova u particiji i konstanta

$$c_{n,i} = \frac{i!}{n!} \times \frac{(n-i)!(i-1)!}{(n-1)!}$$

je odabrana tako da je suma vjerojatnosti jednaka jedan.

Dokaz. Dokaz teorema je moguće pronaći u [2]. ■

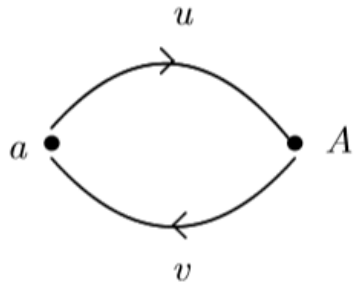
5 Wright - Fisherov model uz prisustvo mutacija

U ovom poglavlju, pretpostavit ćemo i postojanje mutacija u populaciji. Pojmovi korišteni u tekstu ovog poglavlja temeljeni su na [1] i [3].

Neka smo Wright- Fisherov model modificirali na sljedeći način. Pretpostavimo da u svakom vremenskom trenutku jedinka, odmah nakon što je odabrala svog pretka "trpi" mutaciju sljedećeg tipa:

1. alel a postaje alel A sa vjerojatnošću u
2. alel A postaje alel a sa vjerojatnošću v.

Vidimo da se ovdje događaju mutacije sa vjerojatostima u i v, $u, v \in (0,1)$ i mutacije su neovisne za različite pojedince.



Slika 8: Dva tipa mutacija. *Slika je preuzeta iz [1].*

Zanima nas kako mutacija utječe na ponašanje modela?
Iz izraza (4) vidimo da je vjerojatnost da u trenutku $(n+1)$ imamo j alela A kada u trenutku n imamo i alela A jednaka

$$p(i,j) = \binom{2N}{j} p_i^j \left(1-p_i\right)^{2N-j}, \quad i,j = 0,1,\ldots,2N.$$

Napomenimo da je sada vjerojatnost p_i, da u trenutku $(n+1)$ u jednom pokušaju slučajno bude obabran alel tipa A s obzirom da postoji i alela tipa A u trenutku n jednaka

$$p_i = \left(\frac{i}{2N}\right) \cdot (1-v) + \left(\frac{2N-i}{2N}\right) \cdot v. \tag{14}$$

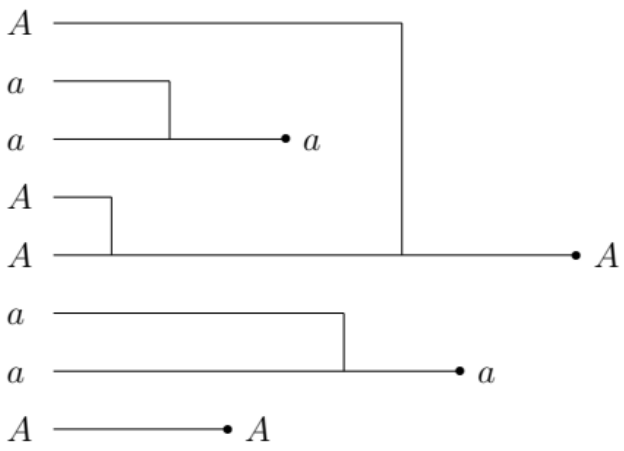
Odnosno, ili je sa vjerojatnošću $\frac{i}{2N}$ izvučen alel A i on ne mutira u alel a ili je sa vjerojatnošću $\frac{2N-i}{2N}$ izvučen alel a i on mutira u alel A.

Neka je X_n, $n \in \mathbb{N}$ slučajna varijabla koja modelira broj alela A u n-toj generaciji. Jedna od posljedica prisutnosti mutacija je da apsorbirajuća stanja 0 i $2N$ nestaju. Prijelazna vjerojatnost u slučaju modela s mutacijama je $p(i,j) > 0, \forall i,j \in S$ jer nema više apsorbirajućih stanja. Mutacije najznačajnije pridonose genetičkoj varijabilnosti unutar populacije. Budući da je skup stanja konačan, slijedi da kada broj generacija $n \to \infty$, $P(X_n = i)$ konvergira ka granici $\pi(i)$ što

je jedinstvena stacionarna distribucija ovo Markovljevog lanca, tj. jedinstveno rješenje sustava jednadžbi

$$\sum_i \pi(i)p(i,j) = \pi(j)$$

gdje je $\pi(i) \geq 0$ i $\sum_i \pi(i) = 1$.



Slika 9: Koalescencija uz prisustvo mutacija. *Slika je preuzeta iz [1].*

Stacionarnu distribuciju najlakše je opisati terminima koalescencije. Označimo tada sa u vjerojatnost da su dva alela koalescirala u alel A, sa v vjerojatnost da su dva alela koalescirala u alel a te sa $(1-u-v)$ vjerojatnost da čestice skaču na slučajno odabrano mjesto, odnosno da se te čestice nisu spojile. Pretpostavljamo da su u i v maleni i ignoriramo pojavu dvije mutacije u jednom koraku. Spajanje čestica određuje njihovo stanje, ali isto tako i stanje svih njihovih potomaka. Ukoliko dođe do toga da se sve čestice spoje i prije nego li dođemo do MRCA, onda stanje u vremenu n ne ovisi o početnom rasporedu i nalazi se u ravnoteži. Slučajan raspored određen kretanjem tog procesa do završetka daje stacionarnu distribuciju za Wright-Fisherov model s mutacijama i X_n konvergira stacionarnoj distribuciji za $n \to \infty$.

Napomenimo da se u sljedećim teoremima radi o konvergenicji po distribuciji te da X_∞ ima graničnu distribuciju. Sljedeći teoremi i njihovi dokazi preuzeti su iz [3].

Teorem 5.1. Neka $X_\infty = \lim_{n\to\infty} X_n$. Tada je matematičko očekivanje jednako

$$EX_\infty = 2N\rho = 2N \cdot \frac{u}{u+v}, \tag{15}$$

gdje je $\rho = \frac{u}{u+v}$ vjerojatnost da smo s vremenom prvo naišli na alel A. (Svaka od $2N$ genetskih veza s vremenom nailazi na alel A ili a.)

Dokaz. Pogledajmo najprije očekivanje EX_n. Iz izraza (14) slijedi da

$$EX_{n+1} = (1-v) \cdot EX_n + (2N - EX_n) \cdot u. \tag{16}$$

Stavimo li da je $x = EX_n = EX_{n+1}$ onda imamo da

$$x = (1 - v)x + (2N - x)u. \tag{17}$$

Rješavanjem dobijemo $(v + u)x = 2Nu$, odnosno imamo da je

$$x = \frac{2Nu}{(u + v)}.$$

Kako bi smo vidjeli da EX_n konvergira ka svojoj granici, primjetimo da ukoliko stavimo $x = 2N\rho$ u izraz (17) imamo

$$2N\rho = 2N(1 - v)\rho + 2N(1 - \rho)u. \tag{18}$$

Oduzmemo li izraz (18) od izraza (16) imamo

$$\begin{aligned}
E(X_{n+1} - 2N\rho) &= (1 - u - v)E(X_n - 2N\rho) \\
EX_{n+1} - 2N\rho &= (1 - u - v)EX_n - (1 - u - v)2N\rho \\
EX_{n+1} - (1 - u - v)EX_n &= 2N\rho - (1 - u - v)2N\rho \\
EX_{n+1} - (1 - u - v)EX_n &= 2N\rho - 2N\rho + 2N\rho u + 2N\rho v \\
EX_{n+1} - (1 - u - v)EX_n &= 2N\rho u + 2N\rho v
\end{aligned}$$

Kako je $EX_n = EX_{n+1}$ slijedi da

$$\begin{aligned}
EX_n - (1 - u - v)EX_n &= 2N\rho u + 2N\rho v \\
EX_n(1 - 1 + u + v) &= 2N\rho u + 2N\rho v \\
EX_n(u + v) &= 2N\rho(u + v)/ : (u + v) \\
EX_n &= 2N\rho
\end{aligned}$$

Slijedi da $EX_N \to 2N\rho$ kada $n \to \infty$. ■

Teorem 5.2. Ukoliko je μ vjerojatnost mutacije u jednoj generaciji, onda je vjerojatnost da su dvije jedinke identične po podrijetlu (kada je μ malen, a N velik) približno jednaka

$$\approx \frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} = \frac{1}{1 + 4N\mu}. \tag{19}$$

Dokaz. Mi na svakom koraku možemo imati mutaciju na nekoj od genskih veza. To nam onda predstavlja događaj sa vjerojatnošću $p_1 = 2\mu$. Također, genske veze mogu i koalescirati, tj. spojiti se, a to nam onda predstavlja događaj sa vjerojatnošću $p_2 = \frac{1}{2N}$. Ukoliko mi u obzir uzmemo jedan ciklus, možemo vidjeti da vjerojatnost ρ, vjerojatnost mutacije prije spajanja zadovoljava jednakost

$$\rho = p_1 + (1 - p_1)(1 - p_2)\rho. \tag{20}$$

ukoliko se niti jedan od događaja ne dogodi, počinjemo ponovno. Ukoliko zanemarimo činjenicu da se mutacija i koalescencija dogode na istom koraku mi prethodnu jednadžbu možemo zapisati kao

$$\rho = p_1 + (1 - p_1 - p_2)\rho. \tag{21}$$

Rješavanjem dobivamo

$$\begin{aligned} \rho &= p_1 + (1 - p_1 - p_2)\rho \\ \rho &= p_1 + \rho - p_1\rho - p_2\rho \\ (p_1 + p_2)\rho &= p_1 / : (p_1 + p_2) \\ \rho &= \frac{p_1}{p_1 + p_2}. \end{aligned}$$

Uvrštavanjem vjerojatnosti p_1 i p_2 dobivamo

$$\rho = \frac{2\mu}{2\mu + \frac{1}{2N}}. \tag{22}$$

Kako su dvije jedinke identične po podrijetlu ukoliko se njihove genske veze spoje prije nego li mutacija djeluje na ijednu od veza, a ρ vjerojatnost mutacije onda gledamo jednakost

$$\begin{aligned} 1 - \rho &= 1 - \frac{2\mu}{2\mu + \frac{1}{2N}} \\ 1 - \rho &= \frac{2\mu + \frac{1}{2N} - 2\mu}{2\mu + \frac{1}{2N}} \\ 1 - \rho &= \frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} \end{aligned}$$

Tako smo došli do željenog izraza, a sređivanjem dolazimo do

$$\frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} = \frac{1}{1 + 4N\mu}. \tag{23}$$

■

Teorem 5.3. Neka je $X_\infty = \lim_{n\to\infty} X_n$. Varijanca u Wright - Fisherovu modelu s mutacijama zadovoljava jednakost

$$VarX_\infty = \left(2N + \frac{2N(2N-1)}{1 + 4N(u+v)}\right) \cdot \frac{uv}{(u+v)^2}. \tag{24}$$

Dokaz. Znamo da općenito vrijedi

$$VarX = EX^2 - (EX)^2.$$

Da bi smo izračunali EX_∞^2, započet ćemo sa promatranjem

$$X_\infty = \sum_{i=1}^{2N} \eta_i,$$

gdje je $\eta_i = 1$ ako je i-ta jedinka alel A, dok je inače 0. U teoriji vjerojatnosti η se naziva indikator varijabla. On ukazuje na to je li se neki događaj dogodio ili nije. Ukoliko kvadriramo prethodnu sumu dobivamo

$$X_\infty^2 = \sum_{i=1}^{2N} \sum_{j=1}^{2N} \eta_i \eta_j. \tag{25}$$

Ukoliko odvojimo $2N$ sa uvjetom $i = j$ od $2N(2N-1)$ s $i \neq j$ tada dobivamo

$$E(X_\infty^2) = 2NP(\eta_1 = 1) + 2N(2N-1)P(\eta_1 = 1, \eta_2 = 1). \tag{26}$$

Iz izraza (15), vidimo da je $P(\eta_1 = 1) = \frac{u}{u+v}$.
Korištenjem prethodnog teorema uz $\mu = u + v$ i razmatranjem moućnosti spajanja čestica prije mutacije ili ne, slijedi da

$$P(\eta_1 = 1, \eta_2 = 1) = \frac{1}{1+4N\mu} \frac{u}{u+v} + \frac{4\mu}{1+4N\mu} \left(\frac{u}{u+v}\right)^2. \tag{27}$$

Sada imamo da

$$(EX_\infty)^2 = 4N^2 \left(\frac{u}{u+v}\right)^2 \tag{28}$$

$$= \left(2N + 2N(2N-1)\left\{\frac{1}{1+4N\mu} + \frac{4N\mu}{1+4N\mu}\right\}\right) \left(\frac{u}{u+v}\right)^2. \tag{29}$$

Korištenjem izraza (25), (26) i (27) dobivamo željeni izraz. ■

Za više detalja pogledati [3].

6 Model beskonačnog alela

U ovom poglavlju reći ćemo nešto više o modelu beskonačnog alela. Pojmovi korišteni u tekstu ovog poglavlja preuzeti su iz [2] i [1].

Kao što smo i rekli u ovom poglavlju razmatrat ćemo model beskonačnog alela, modifikaciju Wright - Fisherova modela u kojem umjesto dva tipa alela postoji beskonačno mnogo tipova alela i svaki put kada se dogodi nova mutacja ona donosi novi tip alela. Dakle, pretpostavlja se da ima toliko alela da je svaka mutacija uvijek novi tip koji nikada ranije nije viđen. Motivacija za ovaj model je sljedeća. Kimura (1971) je tvrdio: Ako se gen sastoji od, primjerice 500 nukleotida, onda broj mogućih DNK lanaca predstavlja varijacije sa ponavljanjem od 500 elemenata četvrte klase. Tada za svaki od njih postoji $3 \cdot 500 = 1500$ lanaca do kojih može doći jednom promjenom osnovnog para. Stoga je vjerojatnost da se vrati gde je počeo u dvije mutacije 1/1500, što je vrlo malo (pod pretpostavkom jednake vjerojatnosti za sve zamjene). Dakle, ukupan broj mogućih alela u biti je beskonačan.

Označimo tipove alela sa $0, 1, 2, \ldots$. Prepostavljamo da se populacija sastoji N jedinki te da svaka jedinka započinje sa tipom 0. U svakom vremenskom trenutku svaka jedinka sa vjerojatnošću $1 - \mu$ slučajnim odabirom bira svog pretka i usvaja njegov tip te sa vjerojatnošću μ mutira u novi tip. Sve se jedinke ažuriraju neovisno jedna o drugoj i neovisno o tome kako su se ažurirali u prethodnim vremenima. Prva mutacija u populaciji donosi jedinku tipa 1, druga mutacija jedinku tipa 2, itd. Kako vrijeme prolazi, novi tipovi alela ulaze u populaciju, a stari tipovi alela odumiru. Možemo očekivati da se nakon nekog vremena distribucija broja različitih tipova alela u populaciji gomila oko neke određene distribucije. Pitanje na koje želimo dobiti odgovor je: „Koja je to distribucija?"

Prije nego li pokušamo dati odogovor na ovo pitanje, reći ćemo nešto detajnije o Hoppeovoj urni te se dotaknuti i Ewensove formula uzimanja u uzorak. Sljedeći pojmovi preuzeti su iz [2].

6.1 Hoppeova urna

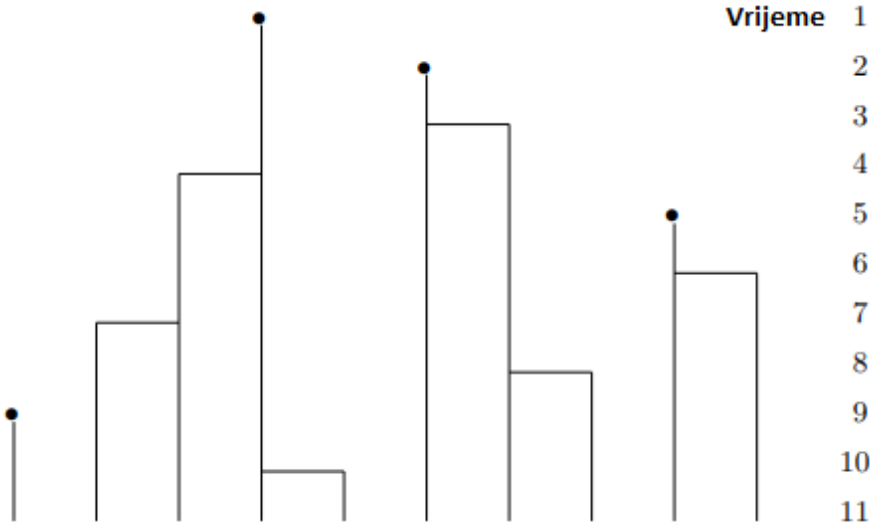
Genealoški proces koji je povezan sa modelom beskonačnog alela je koalescencja sa "ubijanjem". Kada postoji k genskih veza, koalescencija i mutacija se, kao što je i opisano, javljaju u svakom koraku sa vjerojatnošću

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

ali sada se "ubojstvo" jedne od genskih linija događa s vjerojatnošću $k\mu$, gde je μ stopa mutacije po generaciji, jer ako se naiđe na mutaciju zna se genetsko stanje te jedinke i svih njegovih potomaka u uzorku. Ubrzavajući sistem tako što će se raditi po stopi $2N$, stope postaju $\frac{k(k-1)}{2}$ i $\frac{k\theta}{2}$ gde je $\theta = 4N\mu$ i predstavlja Poissonov broj mutacija grane i.

Razmatranje koalescencije s "ubijanjem" unatrag dovodi nas do Hoppeovog modela urne. Neka urna sadrži 1 crnu kuglu (mase θ) i bilo koji broj obojenih kuglica (svaka mase 1). Svaki put, kuglica se slučajnim odabirom bira sa vjerojatnošću proporcionalnoj njenoj masi. Ako je izvučena obojena kuglica, ta kuglica i još jedna iste boje se vraćaju u urnu. Ako je izvučena

crna kuglica, ona se vraća u urnu sa kuglicom nove boje koja ima masu 1. Izbor crne kuglice odgovara novoj mutaciji, a izbor obojene kuglice odgovara koalesenciji. Na sljedećoj slici crna točka označava da je u to vrijeme dodana nova boja.



Slika 10: Realizacija Hoppeove urne. *Slika je preuzeta iz [2].*

Pretpostavimo da gledamo unazad u vremenu, od vremena $(k+1)$ do vremena k u Hoppeovoj urni. Kako imamo k genskih veza, odnosno k obojenih kuglica te jedna crna kuglica pa se mutacija može dogoditi sa vjerojatnošću $\frac{\theta}{\theta+k}$, a koalescencija sa vjerojatnošću $\frac{k}{\theta+k}$. U koalescenciji postoji $(k+1)$ genskih veza koje su svaka izložene mutacijama sa stopom $\frac{k\theta}{2}$, a sudari se javljaju sa stopom $\frac{k(k+1)}{2}$. Budući da su zbog simetrije svi događaji koalescencije imaju jednaku vjerojatnost, slijedi nam sljedeći teorem.

Teorem 6.1.1. Geneološki odnos između k genskih veza u koalescenciji sa ubijanjem može se simulirati pokretanjem Hopeove urne sa k vremenskih koraka.

Ono što je lijepo kod Hoppeova modela urne je to što je to jednostavan postupak, kontroliran parametrom θ i da simulira sve veličine uzoraka odjednom: veličina uzorka n u modelu beskonačnih alela je broj izvlačenja u modelu Hoppeove urne.

Sljedeći rezultat, prema Ewensu bavi se cjelokupnom distribucijom uzorka po modelu beskonačnih alela. Dokaz teorema je moguće pronaći u [2].

Teorem 6.1.2. (**Ewensova formula uzimanja u uzorak**)
Neka je a_i broj tipova alela prisutnih i puta u uzorku veličine n. Kada je stopa mutacije $\theta = 4N\mu$, tada je vjerojatnost da u uzorku koji ima k tipova alela, tip a_1 bude prisutan jednom u uzorku, a_2 dva puta, $\ldots$, a_n n puta jednaka

$$P_{\theta,n}(a_1,\ldots,a_n) = \frac{n!}{\theta(n)}\prod_{j=1}^{n}\frac{(\theta/j)^{a_j}}{a_j!}, \tag{30}$$

gdje je $\theta_{(n)} = \theta(\theta+1)\ldots(\theta+n-1)$.

Za više detalja možete pogledati [1] i [2].

Uzorak male veličine.
Da bi smo lakše shvatili značenje Ewensove formule, razmotrit ćemo male vrijednosti za n.
$n = 2$:
Zapravo gledamo vjerojatnost da u uzorku bude a_1 tipova alela sa jednim i a_2 sa dva alela. Faktor ispred produkta jednak je $\frac{2}{\theta(\theta+1)}$. Dvije su moguće particije: $(a_1, a_2) = (0, 1)$ ili $(2, 0)$. Ako je $(a_1, a_2) = (0, 1)$ to znači da je prisutan jedan tip sa dva alela, a ukoliko je $(a_1, a_2) = (2, 0)$ to znači da su prisutna dva tipa sa po jednim alelom. Produkti u tim slučajevima su $\frac{(\frac{\theta}{2})^1}{1!}$ i $\frac{(\frac{\theta}{1})^2}{2!}$, pa su vjerojatnosti dviju particija $\frac{1}{(\theta+1)}$ i $\frac{\theta}{(\theta+1)}$. Može se zaključiti da je vjerojatnost da su dvije slučajno odabrane jedinke identične (poznato i kao homozigotnost) jednaka $\frac{1}{(\theta+1)}$. To se može zaključiti i direktno na sljedeći način: dvije genske veze se spajaju sa vjerojatnošću $\frac{1}{2N}$ po generaciji i do mutacije dolazi sa vjerojatnoću 2μ. Tada je vjerojatnost da do koalescencije dođe prije mutacije jednaka

$$\frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} = \frac{1}{\theta + 1}. \tag{31}$$

Vratimo se sada na pitanje postavljeno na samom početku ovog poglavlja. To ćemo pitanje razmotriti u granici kada $N \to \infty$, sa

$$\mu = \mu(N) = \frac{\theta}{2N}, \quad \theta > 0,$$

gdje se $\frac{1}{2}\theta$ može smatrati stopom mutacije za cijelu populaciju. U ovom ograničenju, nakon što se vrijeme skalira faktorom N imamo da

$$k \quad \text{genskih veza koalescira sa stopom} \quad \lambda_k = \binom{k}{2} \text{i mutira sa stopom } \frac{1}{2}\theta k.$$

Neka je K_n slučajna varijabla koja broji različite alele koji se nalaze u uzorku veličine n.

Teorem 6.1. (**Watterson**(**1975**))
Za fiksni θ, kada uzorak veličine $n \to \infty$ imamo

$$EK_n \approx \theta \log n \quad \text{i} \quad Var(K_n) \approx \theta \log n. \tag{32}$$

Pored toga vrijedi i centralni granični teorem, odnosno ako slučajna varijabla X ima standardnu normalnu distribuciju, tada:

$$\frac{K_n - EK_n}{\sqrt{Var(K_n)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Prethodni teorem je preuzet iz [2], gdje je moguće pronaći i njegov dokaz.

7 Moranov model

U ovom poglavlju pozabavit ćemo se sa Moranovim modelom, inačicom Wright - Fisherovog modela. Pojmovi korišteni u tekstu ovog poglavlju preuzeti su iz [1] i [2].

Moranov model predstavio je 1958. godine P.A.P. Moran[7], australijski statističar. Kao što smo već naveli, postoji poveznica između Wright - Fisherovog i Moranovog modela. Wright - Fisherov model u obzir uzima nepreklapajuće generacije. Kod ljudi i nekih ostalih vrsta, primjerice vinske mušice to nije slučaj, generacije nisu usklađene. U slučaju takvih vrsta prikladniji je Moranov model, model preklapajućih generacija u kojem se samo jedan alel mijenja tijekom vremena. Model počiva na nekim pretpostavkama:

1. Veličina populacije je konstantna. Da bi se olakšala usporedba s Wright - Fisherovim modelom, pretpostavit ćemo da postoji $2N$ haploidnih jedinki.

2. Svaka jedinka se mijenja po stopi 1.

3. Kada se događa "rođenje" nove jedinke, slučajno se odabire jedna od postojećih $2N$ jedinki(uključujući i odabranu jedinku) da bude roditelj novoj jedinki.

Kao i Wright - Fisherov model, Moranov model opisuje populaciju koja se neutralno razvija i u kojoj su dugovječnost i plodnost jedinke neovisni o genotipu.

Pretpostavimo da su u Moranovom modelu slučajno odabrane dvije jedinke, odnosno dva alela. Odabrani aleli mogu biti istog ili različitog tipa, ali jedan od ta dva alela odabran je da se dalje reproducira, a jedan da nestane, tj. umre. Primjerice, neka smo slučajno odabrali dva alela, A i a. Tada su mogući sljedeći scenariji:

- Alel A je odabran za reprodukciju, a alel a za umiranje.
- Alel a je odabran za reprodukciju, a alel A za umiranje.
- Alel A je odabran za reprodukciju i alel A za umiranje.
- Alel a je odabran za reprodukciju i alel a za umiranje.

Neka je X_t slučajna varijabla kojom modeliramo broj alela A u vremenu t te neka je skup stanja jednak $S = \{0, 1, \ldots, 2N\}$. Kao što smo u ranijim poglavljima rekli, ukoliko u trenutku t postoji i alela A, tada u trenutku t postoji $2N - i$ alela a. Tada ukoliko je proizvoljan alel a zamjenjen alelom A imamo

$$b_i = (2N - i) \cdot \frac{i}{2N}. \tag{33}$$

U izrazu (33) b_i predstavlja vjerojatnost jediničnog povećanja ($i \longrightarrow i + 1$) broja alela tipa A. Odnosno, b_i je vjerojatnost da je odabran alel a i da on u sljedeću generaciju ulazi kao alel A. Isto tako, ukoliko je proizvoljan alel A zamenjen alelom a imamo

[7] https://en.wikipedia.org/wiki/P.A.P.Moran

$$d_i = i \cdot \left(\frac{2N - i}{2N} \right). \tag{34}$$

Slično kao i u izrazu (33), u izrazu (34) d_i predstavlja vjerojatnost jediničnog smanjenja ($i \longrightarrow i - 1$) broja alela tipa A. Odnosno, to je vjerojatnost da je odabran alel A i da on u sljedeću generaciju ne ulazi kao alel A. U ovom slučaju se broj alela A smanjuje jer alel A u sljedeću generaciju ulazi kao alel a. Isto tako, budući da postoje $2N - i$ alela a u trenutku t, vjerojatnost da je jedno od njih zamijenjeno novim alelom je samo $\frac{2N-i}{2N}$.

Možemo primjetiti da su ova dva izraza jednaka. To pokazuje da se broj alela A povećava sa istom stopom sa kojom se i smanjuje.

Sličnost između Moranovog modela i Wright- Fisherovog modela je i ta da oba imaju isto vrijeme apsorbcije. Dakle, i u ovom poglavlju vrijeme apsorbcije možemo definirati na sljedeći način

$$\tau = min\{n : X_n = 0 \quad \text{ili} \quad X_n = 2N\}, \tag{35}$$

odnosno to je trenutak u kojem populacija sadržava sve alele tipa a ili sve alele tipa A.

Sljedeći teorem i njegov dokaz preuzeti su iz [2].

Teorem 7.1. Neka je $u(i) = P(X_\tau = 2N | X_0 = i)$ vjerojatnost da je alel A fiksiran u populaciji veličine $2N$ kada u početku postoji i alela tipa A. U Moranovom modelu vjerojatnost apsorbcije u svim stanjima A jednaka je

$$u(i) = \frac{i}{2N}.$$

Dokaz. Kako je skup stanja S konačan i X_t je ograničen između 0 i $2N$ za sve $t \geq 0$, identitet $E[X_t | X_0 = i] = i$ vrijedi i kada vrijeme $t \geq 0$ zamijenimo sa vremenom apsorbcije τ. Stoga

$$i = E[X_\tau | X_0 = i] = 2N \cdot P(X_\tau = 2N | X_0 = i) + 0 \cdot P(X_\tau = 0 | X_0 = i) = 2N \cdot u(i) \tag{36}$$

Tada dijeljenjem prvog i zadnjeg izraza sa $2N$ dobivamo

$$u(i) = \frac{i}{2N}.$$

■

Neka je uvedena sljedeća oznaka

$$\tilde{E}_i \tau = E_i(\tau | T_{2N} < T_0).$$

pri čemu je prvi T_{2N} trenutak kada su svi aleli tipa A. Tada možemo formulirati sjedeći teorem.

Teorem 7.2. Neka je $p = \frac{i}{2N}$. U Moranovom modelu je očekivano vrijeme apsorbcije kada su svi aleli tipa A, jednako

$$\tilde{E}_i \tau \approx -\frac{2N(1-p)}{p} log(1-p).$$

Dokaz. Dokaz teorema je moguće pronaći u [2]. ■

Literatura

[1] Avena L., da Costa C., den Hollander F., *Stochastic models for genetic evolution*, Mathematical Institute, Leiden University, 2019.

[2] Durrett R., *Probability Models for DNA Sequence Evolution*, 2008.

[3] Durrett R., *Probability Models for DNA Sequence Evolution*, 2002.

[4] Ewens J.W., *Mathematical Population Genetics*, Springer, 2004.

[5] Golub M.,*Genetski algoritam*, 2010., Fakultet elektrotehnike i računarstva Zagreb, http://www.zemris.fer.hr/ golub/ga/ga_skripta1.pdf

[6] Hein J., Schierup M., Wiuf C., *Gene Genealogies, Variation and Evolution*, Oxford, 2005.

[7] Ježić M., *Naslijeđe DNA*, Hrvatsko rodoslovno društvo "Pavao Ritter Vitezović", http://www.rodoslovlje.hr/istaknuta-vijest/naslijede-dna

[8] Pavlica M., *Mrežni udžbenik iz genetike*, Prirodoslovno-matematički fakultet Zagreb, http://www.genetika.biol.pmf.unizg.hr

[9] Praktikum iz Biologije 1, *Geni i nasljeđivanje 1*, Prehrambeno biotehnološki fakultet Zagreb, 2014.-2015.

[10] Škarić-Jurić T., *Uvod u populacijsku genetiku*, Filozofski fakultet Sveučilišta u Zagrebu, 2009./2010.

[11] Šuvak N., *Hardy-Weinbergov model ravnoteže*, Osječki matematički list 5, 2005., 91–99.

[12] Z. Vondraček, *Slučajni procesi*, PMF-Matematički odsjek, Zagreb, 2010.

[13] Z. Vondraček, *Markovljevi lanci*, PMF-Matematički odsjek, Zagreb, 2008.

[14] https://www.nature.com/scitable/topicpage/genetic-drift-and-effective-population-size-772523/

[15] https://www.enciklopedija.hr/natuknica.aspx?id=18721

[16] https://hr.wikipedia.org/wiki/Evolucija

[17] https://bs.wikipedia.org/wiki/Selekcija

Sažetak

U ovom diplomskom radu obrađeni su neki od stohastičkih modela u genetici. Daleko najvažniji i najpoznatiji od svih modela je Wright - Fisherov model. U radu su najprije objašnjene neke od osnovih pojmova genetike. Wright - Fisherov model detaljno je objašnjen bez prisustva i uz prisustva mutacija. Uz njega su objašnjeni i model beskonačnog alela te Moranov model.

Summary

In this paper some of the stochastic models in genetics are explained. The most important and best known of all models is the Wright-Fisher model. In this paper we first explained some of the basic concepts of genetics. The Wright-Fisher model is explained without the presence of mutation and with the presence of mutations. The model of the infinite allele and Moran's model are also explained.

Životopis

Rođena sam 30. lipnja 1996. godine u Bjelovaru. 2003. godine upisujem Osnovnu školu Ivana Nepomuka Jemeršića u Grubišnom Polju, a 2011. godine "Opću gimnaziju" u Srednjoj školi Bartola Kašića u Grubišnom Polju. Tijekom srednjoškolskog školovanja sudjelovala sam na brojnim natjecanjima, među kojima je i Državno natjecanje iz Crvenog križa. Po narodnosti sam Čehinja te aktivno govorim Češki jezik. 2015. godine upisujem preddiplomski studij matematike na Odjelu za matematiku u Osijeku. Dvije godine za redom održavala sam stručno predavanje i radionice učenicima i nastavnicima Srednje škole Bartola Kašića Grubišno Polje. Akademske godine 2018./ 2019. upisujem diplomski studij na istom fakultetu, smjer Financijska matematika i statistika. Tijekom posljednjeg semestra završne godine pohađala sam Pedagoško psihološko didaktičko metodičku izobrazbu na Filozofskom fakultetu u Osijeku te stekla pravo predavanja u osnovnim i srednjim školama.