Matematičko modeliranje u biologiji

Vlahović, Renata

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Renata Vlahović

MATEMATIČKO MODELIRANJE U BIOLOGIJI

Diplomski rad

Voditelj rada:
prof. dr. sc. Željka Milin Šipuš

Suvoditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, rujan 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Draga mentorice, prof. dr. sc. Željka Milin Šipuš, neizmjerno Vam hvala na podršci i ukazanom povjerenju tijekom izrade diplomskog rada, kao i svim godinama mog visokoškolskog obrazovanja u kojima ste uspjeli u meni osnažiti ljubav prema matematici.

Dragi mentore, doc. dr. sc. Pavle Goldstein, od srca Vam se zahvaljujem na suradnji i bezuvjetnoj pomoći u izradi ovog diplomskog rada. Nemoguće je oduprijeti se Vašoj ambicioznosti, upornosti, optimizmu te pozitivnom duhu koji me motivirao tijekom pisanja ovog rada.

Dragi mentori, još jednom veliko Vam hvala!

Hvala mom bratu i prijateljima!

I na kraju, najveće hvala mojim roditeljima na bezuvjetnoj podršci i ljubavi koji su mi pružili tijekom mog obrazovanja!

Sadržaj

Uvod

Poznata izreka velikog njemačkog matematičara Carla Friedricha Gaussa, koji je poznat po svom širokom doprinosu u matematici, fizici i astronomiji, glasi:

"*Matematika je kraljica svih znanosti*".

Veliku ulogu matematika danas ima na području biologije. Matematika, biologija i informatika temelj su bioinformatike. Bioinformatika je interdisciplinarna znanost koja se, između ostalog, bavi analizom bioloških nizova. Na području genetike ona je usmjerena na proučavanje mutacija genoma, dok se u strukturalnoj biologiji bavi analizom DNA, RNA i proteina.

Jedan od glavnih problema kojim se danas bavi bioformatika je traženje varijanti nekog enzima u biološkim organizmima. Cilj ovog diplomskog rada je opisati jednu od metoda traženja varijanti nekog enzima u organizmu poznatu pod imenom *position-specific scoring matrix* ili, kraće, PSSM. S matematičkog gledišta, cilj je povezati navedenu metodu s matematičkim konceptima na kojima se ona temelji.

Diplomski rad podijeljen je na pet poglavlja. U prvom poglavlju objašnjene su osnovne informacije o proteinima, uključujući njihovu građu, strukturu, funkciju te evoluciju. Drugo poglavlje sadrži osnovne pojmove iz vjerojatnosti koji su korišteni u pisanju ovog rada. Osim toga, u poglavlju su opisana dva poznata teorema, centralni granični teorem i Erdös – Rényijev teorem, koji su usko vezani za proučavanje proteinskih nizova. Treće poglavlje osvrće se na jednu poznatu tehniku u bionformatici, a to je poravnanje bioloških nizova te njegov *score*. S matematičkog gledišta zanimljivo je promatrati kako su ti *score*-ovi poravnanja distribuirani. U četvrtom poglavlju opisana je već navedena PSSM metoda za traženje varijanti nekog enzima u nekom organizmu. Varijantu nekog enzima tražimo na temelju već postojećeg skupa varijanti tog enzima koji nazivamo *motiv*. Problem traženja motiva u nekom organizmu poznat je pod imenom *motif scanning*. U tom poglavlju opisani su algoritmi za računanje *score* poravnanja između motiva i proteinskog niza koji promatramo. Na kraju poglavlja naglasak je stavljen na distribuciju maksimalnih *score*-ova. Posljednje poglavlje ovog rada usmjereno je upostavljanje veze između matematike i biologije u osnovnoj i srednjoj školi. Izložene su neke od aktivnosti koje za cilj imaju povezati nastavni sadržaj iz matematike i biologije.

Poglavlje 1

Proteini

Najvažnije tvari u ljudskom organizmu, uz vodu, su *proteini* ili *bjelančevine* (eng. *proteins*). Proteini su izvor tvari za izgradnju mišića, krvi, kože, kose, noktiju i unutarnjih organa, uključujući srce i mozak te su najvažniji čimbenik u rastu i razvoju svih tjelesnih tkiva. Oni su sastavni dio svake stanice, što ih čini osnovom života na Zemlji. Proteini su odgovorni za većinu strukture i biokemijskih atkivnosti svakog živog organizma.

1.1 Građa proteina

Proteini su molekule građene od 20 različitih aminokiselina koje su povezane peptidnom vezom u dugi lanac. Njihov oblik možemo zamisliti kao karike povezane u lanac, pri čemu svaka od karika predstavlja jednu aminokiselinu. Redoslijed i broj tih karika određuju specifične osobine svakog proteina. Promijenimo li samo jednu kariku u tom lancu, tj. neku aminokiselinu zamijenimo drugom, možda ćemo dobiti novi protein. Niz aminokiselina u proteinu određuje njegovu strukturu i funkciju. Proteini se sastoje od stotina ili čak tisuća aminokiselina, a skup proteina čine *proteom* u nekom organizmu.

Aminokiselina (eng. *amino acid*), u biološkom smislu, je osnovna građevna jedinica svakog proteina. U kemijskom smislu, aminokiselina je molekula koja sadrži amino (-NH2) i karboksilnu skupinu (-COOH). Postoje 20 standardnih aminokiselina. Aminokiseline često označavmo njihovim kraticama u obliku jednog ili tri tiskana slova. Nazivi i kratice standardnih aminokiselina nalaze se u tablici 1.1.

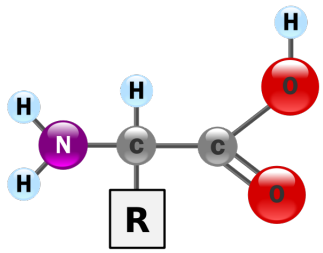
U ljudskom organizmu nalazi se svih 20 aminokiselina, od koji se njih deset mogu izgraditi u samom organizmu, dok je preostalih deset nužno unijeti u organizam kroz prehranu. S obzirom na tu činjenicu, aminokiseline možemo podijeliti u dvije skupine: esencijalne, koje ljudski organizam nije u stanju sintetizirati, a nužne su za funkcioniranje čovjeka, te neesencijalne. Esencijalnim aminokiselinama pripadaju: R, H, L, I, K, M, F, T,

W i V, a neesecijalnim: A, N, D, C, G, Q, E, P, S i Y. Uočimo kako je ova podjela aminokiselina karakteristična za ljudski organizam. U nekom drugom organizmu ova podjela na esencijalne i neesencijalne aminokiseline je drugačija.

Ime	**Kratica**	**Ime**	**Kratica**
Alanin	A (Ala)	Izoleucin	I (Ile)
Arginin	R (Arg)	Leucin	L (Leu)
Asparagin	N (Asn)	Lizin	K (Lys)
Asparaginska kiselina	D (Asp)	Metionin	M (Met)
Cistein	C (Cys)	Prolin	P (Pro)
Fenilalanin	F (Phe)	Serin	S (Ser)
Glicin	G (Gly)	Tirozin	Y (Tyr)
Glutamin	Q (Gln)	Treonin	T (Thr)
Glutaminska kiselina	E (Glu)	Triptofan	W (Trp)
Histidin	H (His)	Valin	V (Val)

Tablica 1.1: Nazivi i kratice standardnih aminokiselina

Osim amino i karboksilne skupine, svaka aminokiselina sadrži određenu R-skupinu. R-skupine sadrže karakterističke osobine pojedinih aminokiselina.

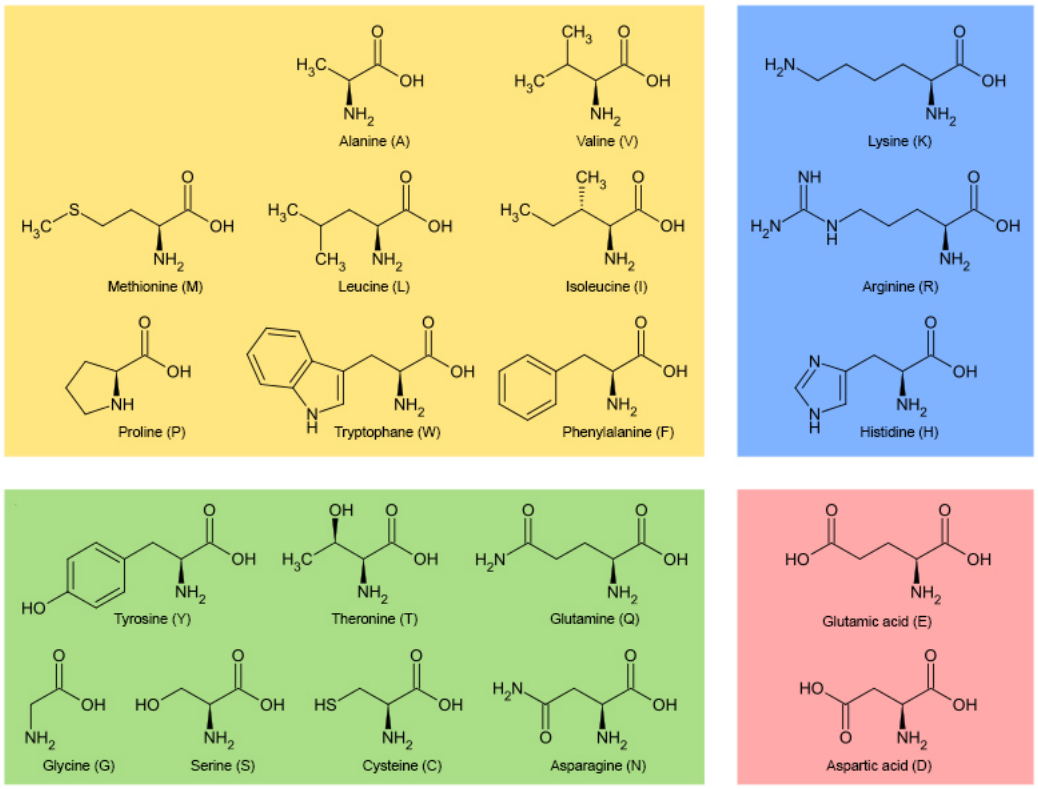


Slika 1.1: Opća struktura aminokiseline

S obzirom na R-skupinu aminokiseline dijelimo na četiri skupine:

1. Aminokiseline nepolarne R-skupine: A, V, L, I, P, F, W i M
2. Aminokiseline polarne R-skupine: G, S, T, C, Y, N i Q
3. Aminokiseline s kiselom R-skupinom: D i E
4. Aminokiseline s bazičnom R-skupinom: K, R i H.

Grafički prikaz podjele aminokiselina prikazan je na slici 1.2.



Slika 1.2: Podjela aminokiselina s obzirom na R-skupinu

1.2 Struktura proteina

Govoreći o građi proteina već smo rekli kako niz aminokiselina u proteinu određuje njegovu strukuru. Struktura proteina je veoma složena te razlikujemo nekoliko nivoa: primarna, sekundarna, tercijarna i kvartalna struktura.

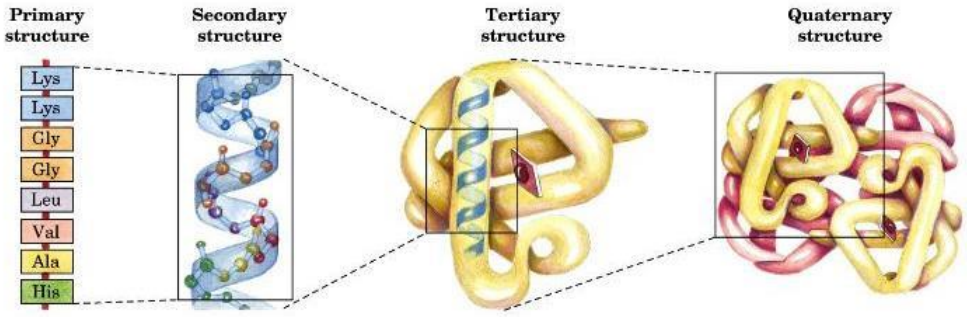
Primarna struktura proteina (eng. *primary structure*)je zapravo *aminokiselinskih niz* (eng. *amino acid sequence*). Niz sastavljen od više aminokiselina povezanih peptidnim vezama nazivamo još i polipeptidni lanac. Ova struktura proteina određena je genima. Svaki polipeptidni lanac ima određenu prostornu strukturu: sekundarnu, tercijarnu ili kvaternu.

Sekundarna struktura proteina (eng. *secondary structure*) je prostorna organizacija uvjetovana interakcijama, najćešće vodikovim vezama, među atomima aminokiselina bliskih u primarnoj strukturi. U obzir se uzima samo okosnica dok se bočni ogranci zanemaruju. Elementi sekundarne strukture su α-zavojnica (eng. *α-helix*), β-ploča (eng. *β-sheet*) i okret (eng. *turn*).

Tercijarna struktura proteina (eng. *tertiary structure*) je prostorna organizacija koja je

nastala interakcijama među aminokiselinama koje nisu blizu u primarnoj strukturi. Postizanjem tercijarne strukture protein postaje biološki aktivan te se on sada u punom smislu može nazvati proteinom. Informacija o teracijarnoj strukturi proteina važna je za razumijevanje funkcije i evolucije proteina.

Kvaterna struktura proteina (eng. *quaternary structure*) je proteinski kompleks koji je nastao udruživanjem više proteina u veće agregate.



Slika 1.3: Shematski prikaz strukture proteina

1.3 Funkcija proteina

Ovisno o svojoj građi, proteini provode čitav niz različitih aktivnosti u ljudskom organizmu. Prva i osnovna funkcija proteina je sudjelovanje u procesu rasta i razvoja. Za bilo koji dio našeg tijela koji prolazi kroz proces rasta ili regeneracije, stvaraju se nove tjelesne stanice, koje trebaju proteine za svoju izgradnju i uspostavljanje odgovarajuće funkcije.

Druga, također velika, funkcija proteina je nadomještanje oštećenih i odumrlih stanica. Stanice koje trebaju uobičajeni nadomjestak jesu između ostalih: stanice krvi, bubrega, jetre, mišića, te naravno stanice kose, nokti, zubi i kosti.

Također, proteini imaju ulogu enzima (molekule koje ubrzavaju biokemijske procese i zaslužne su za ovakav oblik života kakav mi poznajemo) te su oni potrebni tijelu kako bi ono moglo stvoriti hormone (molekule koje omogućuju komunikaciju i usklađivanje biokemijskih procesa između različitih tkiva i organa) i protutijela (molekule koje su proizvod imunološkog sustava organizma i odgovorne su za obranu od stranih tvari, bakterija i virusa).

Jedan od najvažnijih proteina u našem tijelu je hemoglobin - tvar koja prenosi kisik našim tijelom i omogućuje nam odvijanje procesa disanja u svim stanicama u kojima se taj ciklus odvija.

1.4 Evolucija proteina

Istraživanje evolucije proteina vrlo je važno kako bi se bolje razumjela struktura i funkcija proteina što je od velikog interesa zbog važnosti i raznolikosti proteina u ljudskom organizmu. Današnji proteini nastaju kao posljedica evolucije postojećih proteina. Skup proteina koji potječu od istog pretka nazivamo *familija proteina*. Kada govorimo o evoluciji proteina, ustvari mislimo na mutaciju proteina. Postoje tri različita oblika mutacije proteina:

1. *Supstitucija* (eng. *substitution*) - zamjena jedne aminokiseline drugom
2. *Insercija* (eng. *insertion*) - ubacivanje jedne ili više aminokiselina u niz
3. *Delecija* (eng. *deletion*) - izostavljanje jedne ili više aminokiselina iz niza.

Pokažimo na sljedećem primjeru svaku od mutacija:

Primjer 1.1. *Neka je* `DELFIN` *niz aminokiselina.*

Supstitucija: `DELFAN`*. Aminokiselinu* `I` *u nizu* `DELFIN` *zamijenili smo aminokiselinom* `A`*.*

Insercija: `DELUFIN`*. U niz* `DELFIN` *umetnuli smo aminokiselinu* `U`*.*

Delecija: `DFIN`*. Iz niza* `DELFIN` *izostavili smo aminokiseline* `E` *i* `L`*.*

Predak svih nastalih mutacija je niz `DELFIN` *te stoga kažemo da nizovi* `DELFAN`*,* `DELUFIN` *i* `DFIN` *pripadaju istoj familiji proteina.*

Poglavlje 2

Slučajne varijable i distribucije

2.1 Osnovni pojmovi iz teorije vjerojatnosti

2.1.1 Prostor elementarnih događaja

Promotrimo jednostavni pokus bacanja simetričnog novčića. Pokus izvodimo tako da novčić bacamo uvis iznad neke ravne plohe. Nakon što je novčić pao na plohu, na njegovoj gornjoj strani može se nalaziti pismo (P) ili glava (G). Dakle, mogući ishodi ovog pokusa bacanja simetričnog novčića su P i G. Uz navedene pretpostavke, bacanje novčića je pokus kod kojeg će svaki ishod biti element skupa

$$\Omega = \{G, P\}.$$

Uočimo kako ne znamo unaprijed koji će biti ishod bacanja simetričnog novčića.

Pokus sličan ovome je bacanje simetrične kocke kod koje će svaki ishod biti jedan od elemenata skupa

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

ali zbog dane pretpostavke (simetričnosti kocke) ne možemo unaprijed reći koji će to od elemenata biti.

Bacanje simetričnog novčića i simetrične kocke su tzv. primjeri slučajnih pokusa.

Definicija 2.1. ***Slučajan pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

Kod svakog pokusa osnovno je ustanoviti odnos između uzroka i posljedice. Poznavanje tog odnosa omogućuje definiranje uvjeta pokusa i predviđanje ishoda pri svakom realiziranju tih uvjeta. S obzirom na taj odnos razlikujemo dvije osnovne grupe pokusa: determinističke i slučajne. U ovom diplomskom radu nama će biti zanimljivi slučajni pokusi kod kojih ishodi pokusa nisu jednoznačno određeni uvjetima pokusa.

Definicija 2.2. ***Događaj*** *je rezultat slučajnog pokusa.*

Iz toga lako zaključujemo da su događaji bacanja simetričnog novčića "palo je pismo" i "pala je glava". Isto tako, bacamo li novčić pet puta, "pale su tri glave" je također događaj.

Osnovni polazni objekt u teoriji vjerojatnosti je neprazni skup Ω koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnog pokusa. Skup Ω i njegovi elementi, koje često nazivamo i točke skupa Ω, su osnovni i nedefinirani pojmovi u teoriji vjerojatnosti. Točke ω skupa Ω zvat ćemo **elementarnim događajima**. Uočimo da je događaj ω podskup prostora elementarnih događaja Ω.

Cijeli prostor Ω zovemo **siguran događaj** (on se mora dogoditi u svakom vršenju pokusa). Prazan skup $\varnothing$ je **nemoguć događaj** (on se neće nikada dogoditi).

Definicija 2.3. *Neka je A mogući događaj nekog slučajnog pokusa. Pretpostavimo da smo taj pokus ponovili n puta i da se u tih n ponavljanja događaj A pojavio točno n_A puta. Tada broj n_A zovemo* ***frekvencija*** *događaja A, a broj* $\frac{n_A}{n}$ ***relativna frekvencija*** *događaja A (u danih n ponavljanja pokusa).*

Iz definicije relativne frekvencije očito je da vrijedi

$$0 \leq \frac{n_A}{n} \leq 1.$$

2.1.2 Vjerojatnosni prostor

Neka je Ω prostor elementarnih događaja. Sa $\mathcal{P}(\Omega)$ označimo partitivni skup od Ω.

Definicija 2.4. *Familija $\mathcal{A}$ podskupova od Ω jest* ***algebra skupova*** *(na Ω) ako je:*

1. $\varnothing \in \mathcal{A}$
2. $A \in \mathcal{A} \Longrightarrow A^c \in \mathcal{A}$
3. $A_1, A_2, ..., A_n \in \mathcal{A} \Longrightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A}$

Definicija 2.5. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \in \mathcal{F}(\Omega)$) je* ***$\sigma$-algebra skupova*** *(na Ω) ako je:*

1. $\varnothing \in \mathcal{F}$
2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F}, i \in \mathbb{N} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 2.6. *Neka je $(\Omega, \mathcal{F})$ σ-algebra na skupu Ω. Uređeni par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

Definicija 2.7. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** (na $\mathcal{F}$, na Ω) ako vrijedi:*

1. $\mathbb{P}(A) \geq 0, \quad A \in \mathcal{F}$ *(nenegativnost)*
2. $\mathbb{P}(\Omega) = 0$ *(normiranost)*
3. $A_i \in \mathcal{F}, i \in \mathbb{N}\ i\ A_i \cap A_j = \varnothing\ za\ i \neq j \Longrightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *(σ-aditivnost ili prebrojiva aditivnost)*

Definicija 2.8. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je σ-algebra na skupu Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se **vjerojatnosni prostor**.*

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Elemente σ-algebre $\mathcal{F}$ zovemo **događaji**, a broj $\mathbb{P}(A), A \in \mathcal{F}$ zove se **vjerojatnost događaja** A.

Ako je Ω prebrojiv, onda vjerojatnosni prostor $(\Omega, \mathcal{F}, \mathbb{P})$ nazivamo **diskretni vjerojatnosni prostor**.

Vjerojatnost je osnovni objekt u teoriji vjerojatnosti. Često se umjesto termimom vjeroatnost koristimo termimom **vjerojatnosna mjera**.

2.1.3 Uvjetna vjerojatnost. Nezavisnost

Definicija 2.9. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definirajmo funkciju $\mathbb{P}_A : \mathcal{F} \to [0, 1]$ kao:*

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},\ B \in \mathcal{F}.$$

*$\mathbb{P}_A$ je vjerojatnost na $\mathcal{F}$ i zovemo je **uvjetna vjerojatnost uz uvjet A**. Broj $\mathbb{P}(B|A)$ zovemo **vjerojatnost od B uz uvjet A**.*

Definicija 2.10. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A, B \in \mathcal{F}$. Događaji A i B su nezavisni ako vrijedi*

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad i \quad \mathbb{P}(B|A) = \mathbb{P}(B).$$

Dakle, ukoliko su događaji A i B nezavisni, događaj B neće utjecati na vjerojatnost da se dogodi događaj A, i obratno.

Iz definicije uvjetne vjerojatnosti te definicije nezavisnih događaja A i B slijedi da je

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A),$$

odnosno

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Dakle, događaji A i B su nezavisni ako vrijedi

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Uočimo, uvrstimo li dobivenu činjenicu u definiciju uvjetne vjerojatnosti, dobit ćemo

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$$

iz čega možemo zaključiti da je dovoljno da vrijedi $\mathbb{P}(A|B) = \mathbb{P}(A)$ kako bi događaji A i B bili nezavisni.

Generalizacija dobivene posljedice iskazana je sljedećom definicijom:

Definicija 2.11. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A_i \in \mathcal{F}$, $i \in I$ proizvoljna familija događaja. Događaji $A_i \in \mathcal{F}$ su nezavisni ako vrijedi*

$$\mathbb{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i).$$

2.1.4 Slučajna varijabla. Funkcija distribucije

Definicija 2.12. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor.* ***Slučajna varijabla*** *je proizvoljna realna funkcija definirana na Ω.*

Neka je $X : \Omega \to \mathbb{R}$ slučajna varijabla i $B \subset \mathbb{R}$. Tada je:

$$X^{-1}(B) = \{\omega \in \Omega; X(\omega) \in B\} = \{X \in B\}$$
$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{X \in B\}$$

$\mathbb{P}_X(B)$ zovemo vjerojatnosna mjera inducirana sa X.

Ako je $B = (a, b)$, onda je $X^{-1}(B) = \{a < X < b\}$. Ako je $B = (-\infty, a)$, onda je $X^{-1}(B) = \{X \leq a\}$. Ako je $B = \{a\}$, onda je $X^{-1}(B) = \{X = a\}$.

U teoriji vjerojatnosti postoje dva glavna tipa slučajnih varijabli: diskretne i neprekidne. U nastavku ćemo definirati pojmove vezane uz diskretne slučajne varijable, a od neprekidnih slučajnih varijabli navest ćemo tri primjera.

Definicija 2.13. *Slučajna varijabla X je **diskretna** ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}\{X \in D\} = 1$.*

Iz navedene definicije slijedi, ako je X slučajna varijabla takva da je skup svih vrijednosti od X konačan ili prebrojiv, onda je X diskretna. Odavde slijedi, ako je $(\Omega, \mathcal{F}, \mathbb{P})$ diskretan vjerojatnosni prostor, onda je svaka realna funkcija na Ω diskretna slučajna varijabla.

Diskretne slučajne varijable obično zadajemo tako da zadamo skup $D = \{x_1, x_2, ..., x_n, ...\}$ i brojeve $p_n = \mathbb{P}\{X = x_n\}$, što zapisujemo u obliku tablice

$$X = \begin{pmatrix} x_1 & x_2 & ... & x_n & ... \\ p_1 & p_2 & ... & p_n & ... \end{pmatrix}.$$

Dobivenu tablicu zovemo **distribucija slučajne varijabe** X ili **zakon razdiobe** od X.

Definicija 2.14. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ diskretan vjerojatnosni prostor i $X_1, X_2, ..., X_n$ slučajne varijable na Ω. Kažemo da su $X_1, X_2, ..., X_n$ **nezavisne slučajne varijable** ako za proizvoljne $B_i \subset \mathbb{R}, i = 1, ..., n$ vrijedi*

$$\begin{aligned} \mathbb{P}\{X_1 \in B_1, ..., X_n \in B_n\} &= \mathbb{P}\{\omega \in \Omega; X_1(\omega) \in B_1, ..., X_n(\omega) \in B_n\} \\ &= \mathbb{P}(\bigcap_{i=1}^{n} \{X_i \in B_i\}) \\ &= \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\} \end{aligned}$$

Iz definicije slijedi da su slučajne varijable $X_1, X_2, ..., X_n$ nezavisne ako i samo ako su za proizvoljne $B_i \subset \mathbb{R}, i = 1, ..., n$ događaji $\{X_1 \in B_1\}$,...,$\{X_n \in B_n\}$ nezavisni.

Definicija 2.15. *Neka je X slučajna varijabla na Ω. **Funkcija distribucije od** X je funkcija $F_x : \mathbb{R} \to [0, 1]$ definirana sa*

$$\begin{aligned} F_X(x) &= \mathbb{P}_X((-\infty, x]) = \mathbb{P}\left(X^{-1}(-\infty, x]\right) \\ &= \mathbb{P}\{\omega \in \Omega; X(\omega) \leq x\} = \mathbb{P}\{X \leq x\}, x \in \mathbb{R}. \end{aligned}$$

Ukoliko nam je poznato o kojoj se slučajno varijabli, odnosno njenoj funkciji distribucije, radi često koristimo oznaku $F_X = F$.

Definicija 2.16. *Kažemo da je funkcija distribucije F diskretna ako vrijedi*

$$F(x) = \sum_{x_n \leq x, x_n \in D} p(x_n),\ x \in D.$$

Pokaže se da je slučajna varijabla X diskretna ako i samo ako je funkcija distribucije F_X diskretna.

2.1.5 Matematičko očekivanje i varijanca

Definicija 2.17. *Neka je X diskretna slučajna varijabla i neka je $D = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}$ takav da $\mathbb{P}\{X \in D\} = 1$, te neka je $\mathbb{P}\{x_k\} = p_k$ za svako $k = 1, ..., n$.* ***Matematičko očekivanje od*** *X ili, kraće,* ***očekivanje od*** *X, u oznaci $\mathbb{E}X$ definira se sa*

$$\mathbb{E}X = \sum_{k=1}^{n} x_k p_k.$$

Definicija 2.18. *Neka je X diskretna slučajna varijabla i neka $\mathbb{E}X$ postoji.* ***Varijanca od*** *X definira se sa*

$$VarX = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Standardna devijacija od *X, u oznaci σ, je nenegativan kvadratni korijen iz varijance, tj. $\sigma = \sqrt{VarX}$.*

2.1.6 Konvergencija slučajnih varijabli po distribuciji

Neka je $(X_n, n \in \mathbb{N})$ niz slučajnih varijabli definiran na istom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$.

Definicija 2.19. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira po distribuciji prema slučajnoj varijabli X ako je*

$$\lim_{n\to+\infty} F_{X_n} = F_X(x),\ x \in C(F_X)$$

gdje je $C(F_X)$ skup svih točaka neprekidnosti funkcije F_X.

To označavamo $X_n \xrightarrow{\mathcal{D}} X$.

2.1.7 Primjeri neprekidnih slučajnih varijabli

Uočimo kako se sve prethodne definicije vezane uz slučajne varijable odnose na diskretne slučajne varijable. Napomenuli smo da, osim diskretnih slučajnih varijabli, postoje i neprekidne slučajne varijable, no njih ne ćemo posebno definirati. Za njih također možemo definirati funkciju distribucije te matematičko očekivanje i varijancu.

Ako je X neprekidna slučajna varijabla, onda se funkcija distribucije od X naziva *funkcija gustoće vjerojatnosti od X* ili, kraće, *gustoća od X* u oznaci f_X (ili samo f ako nam je poznata distribucija o kojoj se radi). U ovom diplomskom radu nama će biti zanimljive neprekidne slučajne varijable X koje imaju **normalnu**, **gama** ili **Gumbelovu distribuciju**.

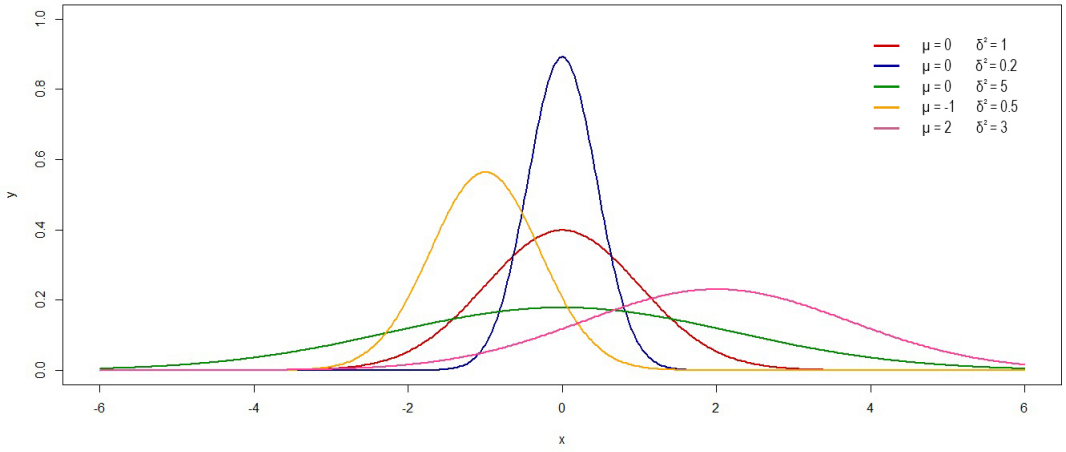
Gaussova ili normalna distribucija

Neka su $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$. Neprekidna slučajna varijabla X ima **normalnu distribuciju s parametrima** m **i** σ^2 ako joj je gustoća f dana sa

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}.$$

To ćemo označavati $X \sim N(\mu, \sigma^2)$.
Očekivanje i varijanca normalne distribucije slučajne varijable X jednake su:

$$\mathbb{E}X = \mu, \quad VarX = \sigma^2. \tag{2.1}$$



Slika 2.1: Funkcija gustoće normalne distribucije s različitim parametrima μ i σ^2

Kažemo da je X **jedinična normalna distribucija** ako je $X \sim N(0, 1)$. Tada je

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \ x \in \mathbb{R}.$$

Funkcija gustoće neprekidne slučajne varijable koja ima normalnu distribuciju još se naziva i **Gaussova funkcija** i označava se s $\varphi(x)$.

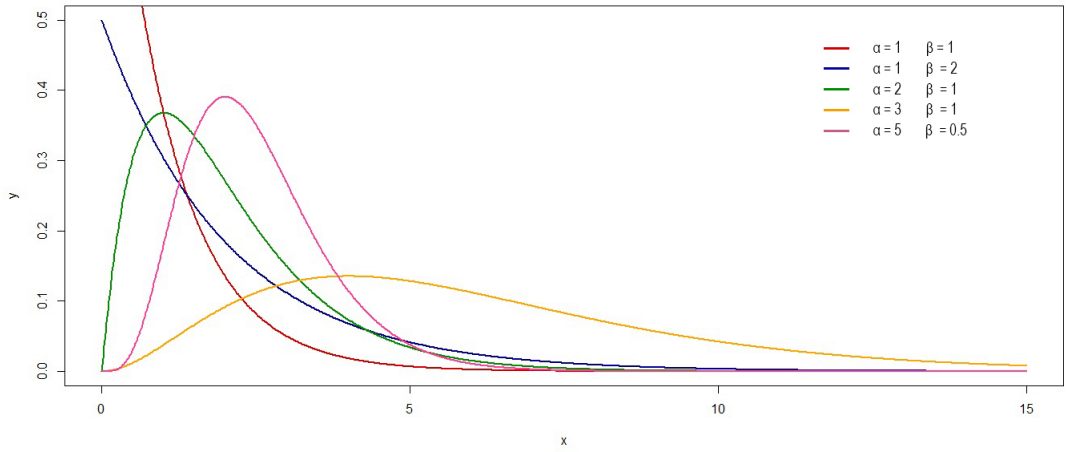
Gama distribucija

Neka je $\alpha > 0, \beta > 0$ i $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt, x > 0$. Neprekidna slučajna varijabla X ima **gama distribuciju s parametrima** α **i** β ako joj je gustoća f dana sa

$$f(n) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Očekivanje i varijanca gama distribucije slučajne varijable X jednake su:

$$\mathbb{E}X = \alpha\beta, \quad VarX = \alpha\beta^2. \tag{2.2}$$



Slika 2.2: Funkcija gustoće gama distribucije s različitim parametrima α i β

Gumbelova distribucija

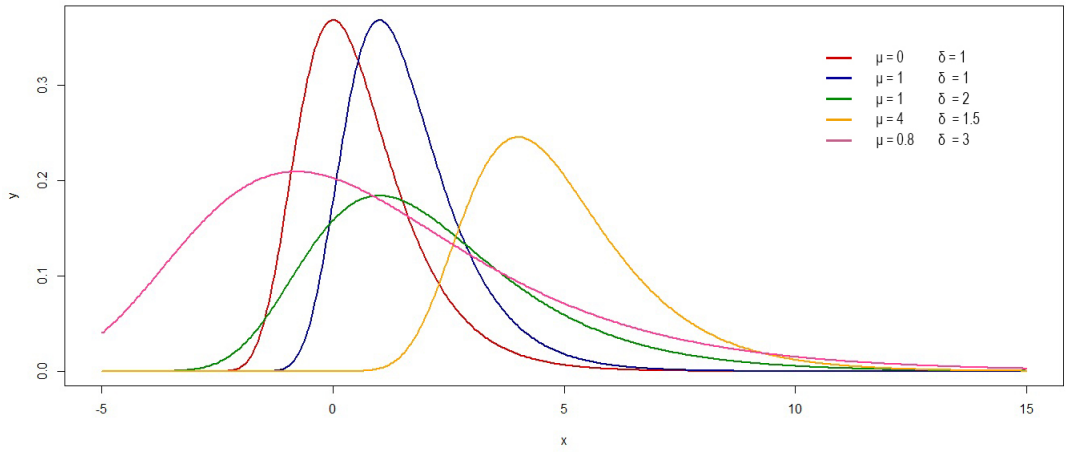
Neka je $\mu, \sigma \in \mathbb{R},\ \sigma > 0$. Neprekidna slučajna varijabla X ima Gumbelovu distribuciju s parametrima μ i σ ako joj je funkcija gustoće f dana s

$$f(x) = \sigma^{-1} exp(-e^{-(x-\mu)/\sigma} - (x-\mu)/\sigma), \qquad x \in \mathbb{R}.$$

Očekivanje i varijanca Gumbel distribucije slučajne varijable X jednake su:

$$\mathbb{E}X = \mu + \sigma\gamma, \quad VarX = \frac{1}{6}\pi^2\sigma^2, \tag{2.3}$$

gdje je $\gamma \sim 0.5772$.



Slika 2.3: Funkcija gustoće Gumbelove distribucije s različitim parametrima μ i σ

2.2 Centralni granični teorem

Prisjetimo se pokusa bacanja simetričnog novčića koji smo opisali na početku poglavlja. Mogući ishodi bacanja novčića su "palo je pismo" i "pala je glava", tj. prostor elementarnih događaja ovog pokusa je skup

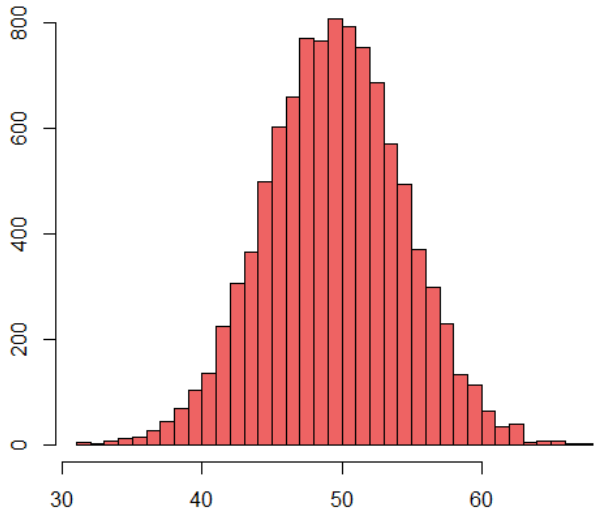
$$\Omega = \{G, P\}$$

gdje P označava događaj da je palo pismo, a G događaj da je pala glava. Pretpostavimo da pokus izvodimo n puta. Sa S_n označimo broj *glava* koje su pale u jednom takvom nizu nastalom ponavljanjem pokusa.

Primjer 2.20. *Pokus bacanja novčića izvodimo 10 puta i kao ishod dobijemo niz PPGPG-GGPGP. Tada je* $S_{10} = 5$.

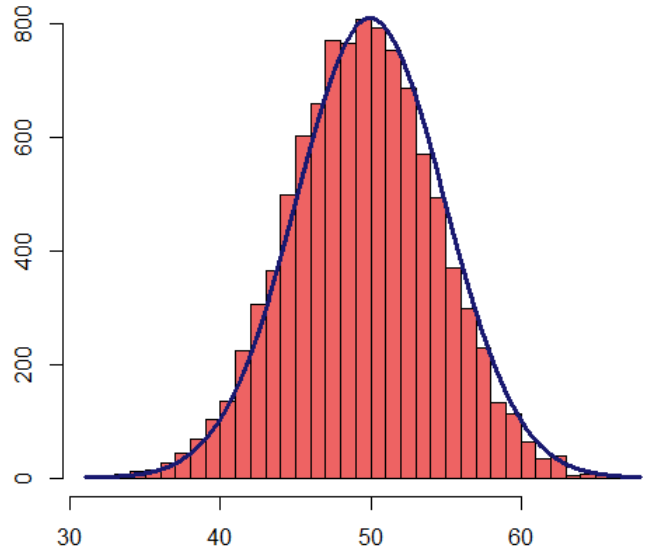
Preostaje nam se pitati kakvu će distribuciju imati varijabla S_n. Na to pitanje odgovor nam daje jedan od najpoznatijih rezultata teorije vjerojatnosti, a to je centralni granični teorem. Centralni granični teorem nam kaže da je distribucija ponavljanja istog pokusa (u našem slučaju bacanje novčića te računanja broja *glava* koje su pale u jednom nizu) konvergira prema normalnoj, tzv. Gaussovoj distribuciji.

Provedimo simulaciju jednog jednostavnog pokusa bacanja simetričnog novčića. Pokus se sastoji od 100 bacanja simetričnog novčića, te taj pokus ponovimo 10000 puta. Distribucija varijable S_n prikazana je na slici:



Slika 2.4: Histogram dobiven simulacijom bacanja simetričnog novčića (duljina nizova: 100, broj simulacijskih nizova: 10000)

U tu činjenicu možemo se uvjeriti i danim histogramom na slici 2.4. Neka je μ srednja vrijednost, a σ^2 uzoračka varijanca. Dobivamo: $\mu = 49.9812$ i $\sigma^2 = 24.26707$. Na sljedećoj slici 2.5 prikazani su dobiveni podaci i funkcija gustoće neprekidne slučajne varijable koja ima normalnu distribuciju s parametrima $\mu = 49.9812$ i $\sigma^2 = 24.26707$ u oznaci $\varphi(\mu, \sigma^2)$.



Slika 2.5: Histogram i funkcija gustoće $\varphi(\mu, \sigma^2)$

Iskažimo spomenuti centralni granični teorem:

Teorem 2.21. ***(Lévy***[1]***)*** *Neka je $(X_n)_{n\in N}$ niz nezavisnih, jednako distribiranih slučajnih varijabli s očekivanjem m i varijancom σ^2, $0 < \sigma^2 < \infty$ i neka je $S_n = \sum_{k=1}^{n} X_n$. Tada vrijedi*

$$\frac{S_n - \mathbb{E}(S_n)}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \quad za \ \ n \to \infty.$$

[1]Paul Pierre Lévy (15.9.1886.-15.12.1971) - francuski matematičkar koji se posebno bavio teorijom vjerojatnosti

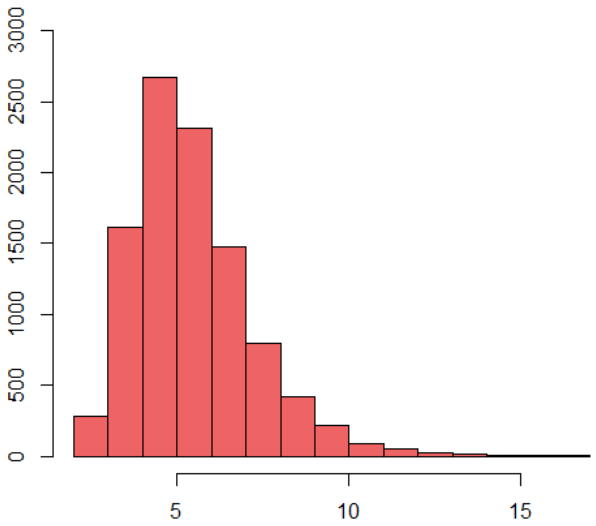
2.3 Erdös – Rényijev teorem

Uočimo da smo u prethodnom poglavlju promatrali broj *glava* u jednom nizu, ne uzimajući u obzir njihov redoslijed. Ono što bismo se također mogli pitati je kolika je duljina najduljeg neprekinutog niza oblika $GGGG\ldots G$.

Pretpostavimo da pokus izvodimo n puta. Sa R_n označimo duljinu najduljeg neprekinutog niza oblika $GGGG\ldots G$. Nazovemo li događaj G *uspjehom*, problem svodimo na traženje najduljeg niza uspjeha. Dakle, R_n nam označava duljinu najvećeg niza uspjeha.

Primjer 2.22. *Pokus bacanja novčića izvodimo 10 puta i kao ishod dobijemo niz PPGPGGGPGP. Tada je najveći niz uspjeha GGG, a duljina najvećeg niza uspjeha* $R_{10} = 3$.

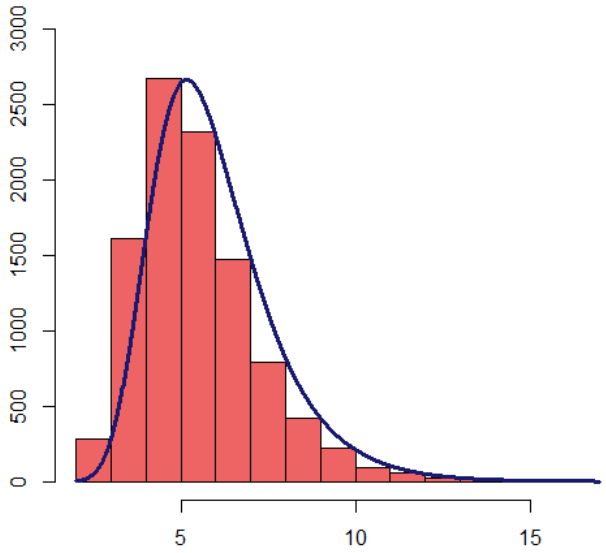
Preostaje nam se pitati kakvu će distribuciju imati varijabla R_n. Vratimo se na simulaciju bacanja simetričnog novčića te pokus koji se sastoji od 100 bacanja simetričnog novčića ponovimo 10000 puta. Distribucija varijable R_n prikazana je na slici:



Slika 2.6: Histogram dobiven simulacijom bacanja simetričnog novčića (duljina nizova: 100, broj simulacijskih nizova: 10000)

Iz prikazanog histograma na slici 2.6 uočavamo kako u ovom slučaju distribucija slučajne varijable R_n nije normalna, već nas ona podsjeća na Gumbelovu distribuciju. Neka je $\bar{X}$ srednja vrijednost, te S^2 uzoračka varijanca. Dobivamo $\bar{X} = 5.9667$ i $S^2 = 3.134305$. Iz 2.3 slijedi da je $\sigma = \sqrt{\frac{6S^2}{\pi^2}}$ i $\mu = \bar{X} - \sigma\gamma$, odnosno $\sigma = 1.380373$ i $\mu = 5.169949$. Na sljedećoj slici 2.7 prikazani su histogram dobivenih podataka i funkcija gustoće ne-

prekidne slučajne varijable koja ima Gumbel distribuciju s parametrima $\sigma = 1.380373$ i $\mu = 5.169949$.



Slika 2.7: Histogram i funkcija gustoće Gumbel distribucije

Pretpostavimo da novčić nije simetričan. Neka je $p = \mathbb{P}(G)$ te neka niz uspjeha R_n ima duljinu m. Tada je vjerojatnost niza uspjeha duljine m jednaka p^m. Uočimo kako su događaji uzastopnog bacanja novčića nezavisni događaji te dobivena vjerojatnost slijedi iz definicije 2.11.:

$$\mathbb{P}(\underbrace{G \cap G \cap \cdots \cap G}_{m}) = \underbrace{\mathbb{P}(G)\mathbb{P}(G)\cdots\mathbb{P}(G)}_{m} = \underbrace{pp\cdots p}_{m} = p^m.$$

Iz pretpostavke da pokus ponavljamo n puta, slijedi da postoji n mogućih nizova uspjeha pa očekivanje da će se broj nizova uspjeha duljine m dogoditi možemo aproksimirati s np^m, tj.

$$\mathbb{E}(\text{broj nizova duljine uspjeha } m) \cong np^m.$$

Ako je najdulji niz uspjeha jedinstven, onda bi trebalo vrijediti da je $\mathbb{E}$(broj nizova duljine uspjeha m) ≈ 1. Iz čega slijedi $1 \approx np^m$, odnosno $1 \approx np^{R_n}$. Time smo došli do aproksimacije duljine najduljeg niza uspjeha R_n, a ona glasi:

$$R_n \approx \log_{1/p} n.$$

Uočimo kako duljina najduljeg niza uspjeha ovisi o duljini niza n.

Primjer 2.23. *Promotrimo kako se mijenja duljina najduljeg niza uspjeha R_n kod simetričnog novčića. Za simetričan novčić vrijedi da je $p = \mathbb{P}(G) = 1/2$. Iz čega slijedi da je $R_n \approx \log_2 n$. Pogledamo li histogram koji prikazuje simulaciju pokusa koji se sastoji od 100 bacanja novčića, dakle $n = 100$, uočit ćemo da dolazimo do istog zaključka: $R_{100} \approx \log_2 100 \approx 6.64$.*

Problem traženja duljine najduljeg niza uspjeha koji smo opisali je ustvari poopćeni problem kojim su se bavili matematičari Erdös[2] i Rényi[3]. Spomenuti matematičari svoj su rezultat prezentirali 1970. godine. Oni su proučavali dva niza jednakih duljina čija su slova nezavisne, jednako distribuirane varijable te su se pitali koja je duljina najvećeg podudarajućeg segmenta među njima. Demonstrirajmo njihov problem na jednostavnom primjeru.

Primjer 2.24. *Dvije osobe bacaju novčić i bilježe svoje ishode. Primjer jednog takvog pokusa:*

$$\begin{aligned} Osoba\ \ I: &\ \ GGPGPPPGGP \\ Osoba\ \ II: &\ \ PGPGPGGGPG \end{aligned}$$

Najveći podudarajući segment između ova dva niza je GPGP, iz čega zakjučujemo da je duljina najvećeg podudarajućeg segmenta jednaka 4.

Naša razmatranja iskažimo formalno sljedećim teoremom:

Teorem 2.25. ***(Erdös – Rényijev teorem)*** *Neka su $A_1, A_2, \ldots, B_1, B_2, \ldots$ nezavisne, jednako distribuirane slučajne varijable i $0 < p = \mathbb{P}(A_i = B_i) < 1$. Definirajmo*

$$R_n = \max\{m : A_{i+k} = B_{i+k} \text{ za sve } k = 1, ..., m, 0 \leq i \leq n - m\}.$$

Tada je

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{R_n}{\log_{1/p} n} = 1\right) = 1.$$

Uočimo, uspjehom u prethodnom primjeru, nazvali smo podudaranje poravnatih slova te smo tražili najveći podudarajući segment. Pogledajmo sljedeći primjer:

[2]Paul Erdös (26.3.1913.-20.9.1996.) - mađarski matematičar koji se bavio kombinatorikom, teorijom grafova, teorijom brojeva, klasičnom analizom, teorijom aproksimacija te teorijom vjerojatnosti

[3]Alfréd Rényi (20.3.1921.-1.2.1970.) - mađarski matematičar koji se bavio kombinatorikom, teorijom grafova i posebice teorijom vjerojatnosti

Primjer 2.26. *Ishodi bacanja novčića dviju osoba dane su nizovima:*

GPPGGPGGPP
PPGGGPPGGG

Računamo li najveći podudarajući segment kao u prošlom primjeru, dobit ćemo da je on duljine 3 i oblika GGP.

GPP GGP GGPP
PPG GGP PGGG

Uočimo kako smo ovim načinom traženja najvećeg podudarajućem segmenta pretpostavili da su nizovi poravnati tako da se ispod svakog slova gornjeg niza nalazi slovo donjeg niza. Ono što još možemo dopustiti je "pomicanje" donjeg niza te traženje najvećeg podudarajućeg segmenta. Dopustimo li pomicanje danih nizova uočit ćemo da je tada duljina najvećeg podudarajućeg segmenta jednaka 5 i on je oblika GPPGG.

GPPGG PGGPP
PPGG GPPGG G

Pretpostavimo da su nizovi duljine n. Neka je p vjerojatnost podudaranja slova u nizovima. S H_n označimo duljinu najvećeg podudarajućeg segmenta te s m označimo njegovu duljinu. Tada je vjerojatnost pojavljivanja najvećeg podudarajućeg segmenta jednaka p^m. Dopustimo li pomake danih nizova, očekivanje da se podudarajući segment duljine m dogoditi možemo aproksimirati s $n^2 p^m$, tj.

$$\mathbb{E}(\text{broj nizova duljine uspjeha } m) \cong n^2 p^m.$$

Pretpostavimo li da je najdulji podudarajući segment jedinstven, trebalo bi vrijediti da je $1 \approx n^2 p^m$, odnosno $1 \approx n^2 p^{H_n}$. Time smo došli do aproksimacije duljine najvećeg podudarajućeg segmenta H_n, a ona glasi:

$$H_n \approx 2 \log_{1/p} n.$$

Dobiveni zaključak možemo i formalno iskazati sljedećim teoremom:

Teorem 2.27. *Neka su $A_1, A_2, \ldots, B_1, B_2, \ldots$ nezavisne, jednako distribuirane slučajne varijable i $0 < p = \mathbb{P}(A_i = B_i) < 1$. Definirajmo*

$$H_n = \max\{m : A_{i+k} = B_{j+k} \text{ za sve } k = 1, ..., m, 0 \leq i, j \leq n - m\}.$$

Tada je

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{H_n}{\log_{1/p} n} = 2\right) = 1.$$

Poglavlje 3

Poravnanje nizova. *Score* poravnanja

Želimo li utvrdili pripadaju li neki proteini istoj familiji, tj. imaju li zajedničku evoluciju, trebamo ih usporediti i utvrditi njihovu sličnost. Osim toga, sličnost dvaju proteinskih nizova može ukazati i na istu funkciju i strukturu tih proteina.

3.1 Poravnanje nizova

Prije samog uspoređivanja nizova, obično proteinske nizove prvo poravnamo. *Poravnanje nizova* (eng. *sequence alignment*) je način uređivanja proteinskih nizova u svrhu identificiranja sličnosti među njima kako bi se utvrdila određena funkcionalna, strukturna i evolucijska povezanost proteina. Utvrdimo li analizom da su proteini "jako slični", možemo pretpostaviti da oni imaju sličnu biokemijsku funkciju i trodimenzionalnu strukturu. Osim toga, utvrdimo li sličnost između proteinskih nizova iz različitih organizama, postoji vjerojatnost da ti nizovi imaju zajedničkog pretka te poravnanje tih nizova može nam ukazati na promjene koje su se mogle dogoditi na tim proteinima tijekom evolucije. Želimo li pronaći poravnanje takvo da je sličnost između proteinskih nizova maksimalna, govorimo o *optimalnom poravnanju*. Kako bi odredili optimalno poravnanje dvaju nizova koristimo se različim algoritmima, a najpoznatiji među njima su *Needleman-Wunsch*[1] i *Smith-Waterman*[2].

Poravnate nizove obično zapisujemo u dva retka, jedan ispod drugoga ili, matematički rečeno, reprezentiramo ih recima matrice. Cilj poravnanja nizova proteina ponekad je pronalaženje što je više moguće poravnatih parova aminokiselina, a ponekad nas više zanima najdulji niz aminokiselina koji se nalazi u oba niza. S obzirom na tu činjenicu razlikujemo dvije vrste poravnanja: *globalno* i *lokalno*. Kako bismo što uspješnije poravnali identične

[1] Saul B. Needleman i Christian D. Wunsch su 1970. godine predstavili prvi algoritam za poravnanje dvaju nizova proteina koji se temelji na dinamičkom programiranju.

[2] Temple F. Smith i Michael S. Waterman svoj su algoritam predstavili 1981. godine.

aminokiseline u proteinima, pri poravnanju, nizovima dodajemo praznine (eng. *gaps*) koje grafički prikazujemo kao '–'. Nizovi koje želimo poravnati ne moraju nužno biti jednake duljine. Osim toga, možemo poravnati samo dva niza, a možemo i njih više. S obzirom na tu činjenicu razlikujemo *dvostruka* i *višestruka* poravnanja.

Na sljedećim nizovima[3] objasnit ćemo razliku između globalnog i lokalnog poravnanja:

Protein primglo	...TLVGSALHPDSRSHPRSLEKSAWRAFKESQ...
Xenopus laevis	
(African clawed frog)	
Protein primglo 1	...DVGQSSALTLSDSRLHPQSLEKSPWREFQC...
Gallus gallus (Chicken)	

3.1.1 Globalno poravnanje

Cilj globalnog poravnanja je poravnati što je više moguće identičnih animokiselina iz oba niza koje uspoređujemo. Nizovi pogodni za globalno poravnanje su nizovi koji su i bez prethodnog poravnanja slični te su otprilike jednake duljine. Najpoznatiji algoritam za globalno poravnanje je *Needleman-Wunsch*.

Primjer globalnog poravnanja navedenih proteinskih nizova:

```
TLVG-S-ALHP-DSRSHPRSLEKSAWRAFKESQ-
D-VGQSSALTLSDSRLHPQSLEKSPWREF---QC
```

3.1.2 Lokalno poravnanje

Cilj lokalnog poravnanja je pronaći što više podudarajućih podnizova danih proteinskih nizova. Lokalno poravnanje je pogodnije za nizove kojima je dio niza ostao očuvan, dok je drugi dio prošao kroz razne evolucijske promjene. Najpoznatiji algoritam za lokalno poravnanje je *Smith-Waterman*.

Primjer lokalnog poravnanja navedenih proteinskih nizova:

```
TLVG--SALHP-DSRSHPRSLEKSAWRAFKESQ
 DVGQSSALTLSDSRLHPQSLEKSPWREFQC
```

[3] Nizovi preuzeti s http://www.uniprot.org/.

3.2 *Score* poravnanja

Kako bismo mogli reći koji proteini su sličniji ili koje poravnanje je bolje, potrebno je izabrati model ocjenjivanja poravnanja koji još nazivamo *score poravnjanja* te označujemo sa S.

Želimo li odrediti *score* **globalnog poravnanja**, jedan od najjednostavnijih modela ocjenjivanja poravnanja je:

$$S = broj\ mjesta\ gdje\ se\ aminokiseline\ podudaraju. \tag{3.1}$$

Uočimo, ovaj model ocjenjivanja je "dobar" ukoliko uspoređujemo sve nizove jednakih duljina. Ukoliko nizovi nisu jednakih duljina, ocjena poravnanja će biti relativna. Pokažimo to na primjeru.

Primjer 3.1. *Zadana su dva para proteinskih nizova:*

ALFD	FLAGTTFI
AGFK	KLHLGFLS

Računamo li score poravnanja prema modelu opisanom u 3.1, score oba poravnanja jednak je 2. Iz toga bi se moglo zaključiti da su oba para ovih nizova jednako slični, što je pogrešno. Lako se izračuna da se u prvom paru nizova 50% aminokiselina podudara, dok se u drugom paru nizova podudara upola manje, tj. 25%.

Iz prethodnog primjera vidimo kako je *score* poravnanja definiran u 3.1 relativan broj. Malo složeniji model ocjenjivanja, a još uvijek daleko najjednostavnji, bi u u obzir uzimao i broj mjesta na kojima se aminokiseline ne podudaraju, te ga možemo definirati na sljedeći način:

$$\begin{aligned} S = &(broj\ mjesta\ gdje\ se\ aminokiseline\ podudaraju)\ - \\ &(broj\ mjesta\ gdje\ se\ aminokiseline\ ne\ podudaraju). \end{aligned} \tag{3.2}$$

Sličnost među nizovima je veća, odnosno poravnanje je bolje, što je veći *score* poravnanja. Računajući *score* poravnanja prema modelu 3.2 dobivano da je *score* prvog poravnanja jednak $2 - 2 = 0$, a *score* drugog poravnanja $2 - 6 = -4$. Iz čega možemo zaključiti da su nizovi iz prvog para sličniji.

Govoreći o **lokalnom poravnanju** rekli smo da je cilj poravnanja pronaći što više podnizova koji se podudaraju. S obzirom na to, najjednostavniji model ocjenjivanja bio bi:

$$S = duljina\ najvećeg\ podudarajućeg\ segmenta. \tag{3.3}$$

Pokažimo to na primjeru:

Primjer 3.2. *Zadana su dva niza koja su lokalno poravnata:*

```
TLVG--SALHP-DSRSHPRSLEKSAWRAFKESQ
DVGQSSALTLSDSRLHPQSLEKSPWREFQC
```

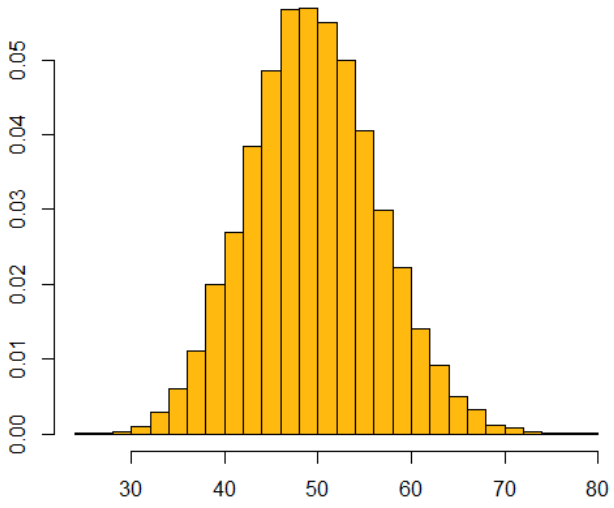
Tada je score lokalnog poravnanja jednak 5, jer najdulji podudarajući segment je `SLEKS`.

3.3 Distribucija *score*-ova poravnanja

Od interesa nam je znati kako su distribuirani ti *score*-ovi poravnanja. Kako bismo to saznali, jedan od načina je provođenje pokusa. U sljedećim simulacijama prikazat ćemo kako su distribuirani *score*-ovi poravnanja definirati s 3.1 i 3.3.

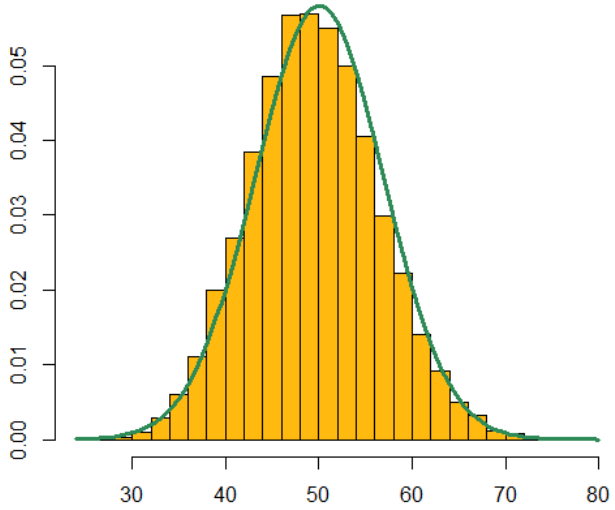
Simulacija 1

Neka je $\mathcal{A} = \{$`A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V`$\}$. U svakom pokusu simuliramo po dva proteinska niza duljine 1000 čiji su elementi iz skupa $\mathcal{A}$. Pokus ponavljamo 10000 puta. Za svaka dva simulirana niza računamo *score* koristeći 3.1. Dakle, tražimo broj mjesta gdje se aminokiseline podudaraju. Distribucija *score*-ova prikazana je na slici:



Slika 3.1: Histogram *score*-ova dvaju proteinskih nizova (duljina proteinskih nizova: 1000, broj pokusa: 10000)

Iz histograma prikazanog na slici 3.1 možemo naslutiti da ovi podaci slijede normalnu distribuciju. Neka μ srednja vrijednost, a σ^2 varijanca tih *score*-ova. Dobivamo: $\mu = 50.0111$ i $\sigma^2 = 47.34331$. Na sljedećoj slici 3.2 prikazani su podaci i funkcija gustoće neprekidne slučajne varijable koja ima normalnu distribuciju s parametrima μ i σ^2 u oznaci $\varphi(\mu, \sigma^2)$.

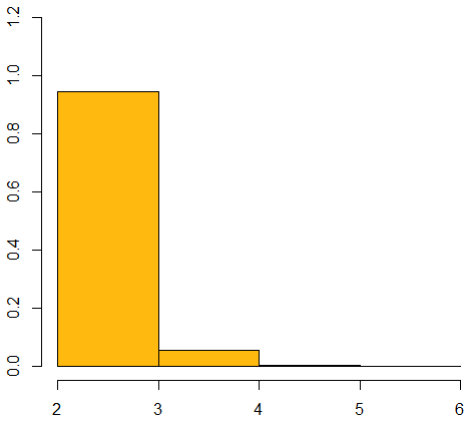


Slika 3.2: Histogram *score*-ova i funkcija gustoće $\varphi(\mu, \sigma^2)$

Prisjetimo li se prethodnog poglavlja, uočit ćemo da nam upravo o tome govori i centralni granični teorem.

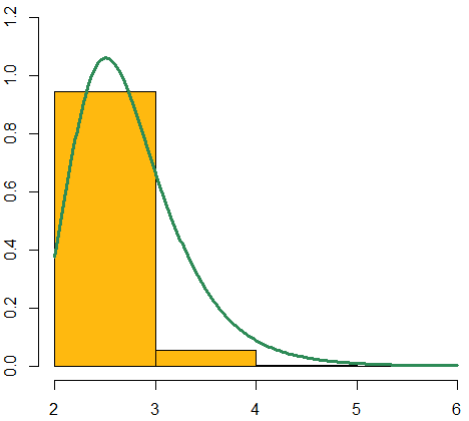
Simulacija 2

Neka je $\mathcal{A} = \{$`A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V`$\}$. U svakom pokusu simuliramo po dva proteinska niza duljine 10000 čiji su elementi iz skupa $\mathcal{A}$. Pokus ponavljamo 10000 puta. Za svaka dva simulirana niza računamo *score* koristeći 3.3. Dakle, tražimo duljinu najvećeg podudarajućeg segmenta dvaju proteinskih nizova. Distribucija *score*-ova prikazana je na slici:



Slika 3.3: Histogram *score*-ova dvaju proteinskih nizova (duljina proteinskih nizova: 1000, broj pokusa: 10000)

Iz histograma prikazanog na slici 3.3 uočavamo kako ovaj puta *score*-ovi ne slijede normalnu distribuciju. Pretpostavimo da dobiveni *score*-ovi slijede Gumbel distribuciju. Neka je $\bar{X}$ srednja vrijednost, te S^2 uzoračka varijanca. Dobivamo $\bar{X} = 2.7548$ i $S^2 = 0.3089079$. Iz 2.3 slijedi da je $\sigma = \sqrt{\frac{6S^2}{\pi^2}}$ i $\mu = \bar{X} - \sigma\gamma$, odnosno $\sigma = 0.4333514$ i $\mu = 2.50467$. Na sljedećoj slici 3.4 prikazani su histogram dobivenih podataka i funkcija gustoće neprekidne slučajne varijable koja ima Gumbel distribuciju s parametrima $\sigma = 0.4333514$ i $\mu = 2.50467$.



Slika 3.4: Histogram *score*-ova i funkcija gustoće Gumbel distribucije

Iako se iz grafičkog prikaza ne vidi najpreciznije da *score*-ovi slijedi Gumbel distribuciju, prisjetimo li se prethodnog poglavlja i Erdös – Rényijevog teorema, uočit ćemo da se radi o istim vrstama distribucija. Traženje duljine najvećeg podudarajućeg segmenta aminokiselina svodi se na traženje duljine najvećeg niza uspjeha. U našem primjeru uspjehom nazivamo podudaranje dviju aminokiselina u nizovima.

Uočimo kako u prethodnim modelima za računanje *score*-ova nismo u obzir uzimali vjerojatnost pojavljivanja aminokiselina u nekom proteinskom nizu. Skrivena pretpostavka provedenih pokusa je da je pojavljivanje svake aminokiseline u proteinskom nizu jednakovjerojatno. Osim toga, u obzir nismo uzeli niti vjerojatnost podudaranja dviju aminokiselina. U prirodi to i nije baš tako jednostavno. Želimo li preciznije ustanoviti sličnost dvaju proteina, trebamo uvesti puno složeniji način određivanja *score*-ova. U obzir trebamo uzeti i vjerojatnost pojavljivanja određene aminokiseline u nizu, ali i vjerojatnost podudaranja dviju aminokiselina. U daljnjem razmatranju promatrat ćemo samo nizove koji ne sadrže praznine.

Jedna od metoda računanja *score*-ova poravnanja je pomoću PSSM matrice o kojoj ćemo nešto više reći u sljedećem poglavlju.

Poglavlje 4

PSSM

4.1 Motiv

Uspoređujući dva niza, uočimo da smo tražili samo mjesta na kojima su bile identične aminokiseline u oba niza. No, ponekad aminokiseline ne moraju biti identične kako bismo rekli da se nizovi podudaraju ili da su "jako slični". Neke aminokiseline su jako slične po svojoj strukturi (na primjer: `V,A,F`), pa jednu aminokiselinu možemo zamijeniti drugom. Isto tako, neke aminokiseline imaju jako sličnu funkciju (na primjer: `V,L,I`), pa opet možemo zamijeniti aminokiseline jednu s drugom.

Pokažimo na primjeru sljedećih nizova[1]:

Zinc finger protein Homo sapiens (Human)	`MDPEQSVKGTKKAEGSPR---VSSSVPYPGSGTAATQESPA`
	`MDP*QS*KGTKKA*GSPR---VSSS*PYPGSGT*A *ES *`
Zinc finger protein Mus musculus (Mouse)	`MDPDQSIKGTKKADGSPR---VSSSAPYPGSGTTAPSESAT`

U srednjem redu ispisali smo sve aminokiseline koje su identične, a između aminokiselina koje su slične stavili smo '*'. Na mjestu gdje aminokiseline nisu identične ili slične ostavili smo prazninu.

Uzmemo li u obzir sličnost aminokiselina te gledamo li na ovaj način koliko se proteini podudaraju, uočit ćemo da su oni "jako slični".

[1] Nizovi preuzeti s http://www.uniprot.org/.

Ovo je primjer u kojem smo tražili sličnost između dva proteinska niza. Isto to možemo učiniti i s više proteinskih nizova.

Tijekom evolucije različitim mutacijama nastale su različite varijante nekih proteina. Promotrimo nekoliko varijanti nekog enzima:

```
VTGSFLDA
LAGDFIDF
LTGSFLDF
VDGDFLDA
VAGDFIDA
```

Promatrajući nizove uočavamo da se na istim mjestima u nizovima javljaju različite aminokiseline. Na primjer, na trećem mjestu u svim nizovima nalazi se aminokiselina `G`. Isto to vrijedi i za peto i za sedmo mjesto gdje se nalazi aminokiselina `F`, odnosno `D`. Na prvom mjestu pojavljuju se aminokiseline `V` i `L`, od čega uočavamo da se aminokiselina `V` ponavlja više puta. Daljnjom analizom uočavamo da se na drugom mjestu mogu naći aminokiseline `A`,`T` i `D`, na četvrtom mjestu aminokiseline `S` i `D`, na šestom mjestu `L` i `I`, a na posljednjem osmom mjestu aminokiseline `A` i `F`. Ono što je bitno uočiti je da se aminokiseline na pojedinim mjestima ne javljaju u jednakom omjeru.

Jedna od glavnih tema u bionformatici je upravo pronalazak varijanti nekih proteina u nekom novom organizmu. Na primjer, neka od pitanja koja bismo mogli postaviti je nalazi li se neka od navedenih varijanti ovog enzima u nekom organizmu ili postoji li još koja nepoznata varijanta ovog enzima u nekom drugom organizmu.

Jedan od načina traženja odgovora na ta pitanja bio bi pronalaženje podnizova u proteinima koji su slični svakom od ovih varijanti enzima. Uzmemo li u obzir da neki proteini (ili dijelovi proteina) mogu imati i do nekoliko stotina aminokiselina, ovakav postupak pretraživanja bio bi poprilično dugotrajan. Kako bismo izbjegli pretraživanje te uspoređivanje dijelova proteina sa svakim od varijanti, ideja je usporediti proteinske nizove istovremeno sa svim varijantama danog enzima.

Skup varijanti nekog proteina, odnosno uzorak varijanti nekog proteina koji nam je poznat nazivamo ***motiv*** (eng. *motif*). Neki biolozi smatraju da je motiv najmanja strukturna jedinica koja opstaje ili nestaje evolucijom. No, motiv je skup nizova od 10-tak do 20-tak aminokiselina na kojem se jasno vide promjene koje su nastale tijekom evolucije proteina.

Na ovaj način smo problem traženja dijelova proteina koji su slični varijantama nekog enzima sveli na traženje dijelova proteina koji su slični danom motivu. Taj pojam traženja motiva u proteinima poznat je još i kao ***motif scanning***.

4.2 *Position-specific scoring matrix* ili PSSM

Uočimo kako ovakav zapis motiva nije prikladan za uspoređivanje i pretraživanje dijelova proteina koji su slični nekom zadanom motivu. Iz tog razloga trebamo pronaći zapis motiva koji je prikladniji za korištenje. Oblik u kojem se motiv zapisuje nazivamo *profil motiva*.

Zbog jednostavnosti zapisa, konstruirajmo *profil motiva* koji sadrži samo osnovne četiri aminokiseline: `A`,`C`,`G` i `T`.

Primjer takvog motiva:

```
TGTCGA
TGTAAA
AGTCTA
GGTGTA
CGATAA
CGATGA
AGAGCA
AGTTCA
```

Kao i u prošlom primjeru možemo uočiti kako se na određenim mjestima u nizovima ne pojavljuju uvijek identične aminokiseline, niti se one pojavljuju u jednakom omjeru. S obzirom na tu činjenicu, ideja je da *profil motiva* sadrži učestalost pojavljivanja aminokiselina na određenim mjestima. Jedan od načina računanja učestalosti aminokiselina je računanje frekvencija njihovih pojavljivanja na odredeđnim mjestima u nizovima.

Matrica frekvencija:

	A	C	G	T	
1	3	2	1	2	8
2	0	0	8	0	8
3	3	0	0	5	8
4	1	2	2	3	8
5	2	2	2	2	8
6	8	0	0	0	8

Matricu frekvencija konstruirali smo tako da i-ti stupac matrice označuje broj određene aminokiseline na j-tom mjestu u motivu koje smo prikazali recima matrice. Ono što nas više zanima je vjerojatnost pojavljivanja određene aminokiseline na nekom mjestu u motivu. Za to su nam potrebne relativne frekvencije pojavljivanja aminokiselina na određenim mjestima u motivu.

Matrica relativnih frekvencija:

	A	C	G	T	
1	0.375	0.25	0.125	0.25	1
2	0	0	1	0	1
3	0.375	0	0	0.625	1
4	0.125	0.25	0.25	0.375	1
5	0.25	0.25	0.25	0.25	1
6	1	0	0	0	1

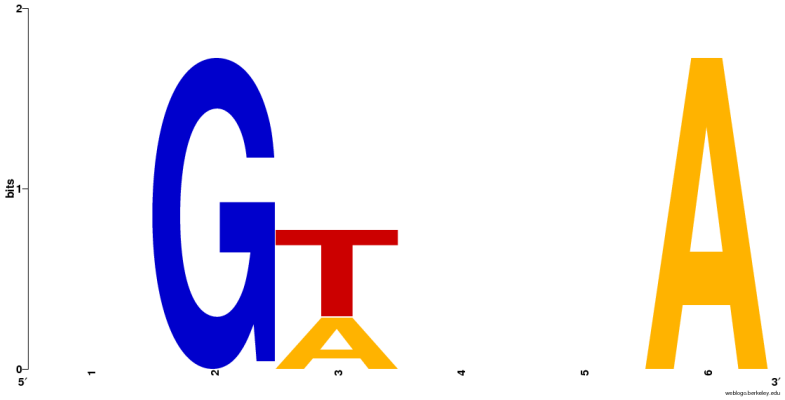
Matrica relativnih frekvencija je ustvari matrica vjerojatnosti pojavljivanja određene aminokiseline na određenom mjestu u motivu. S obzirom na tu činjenicu ovako konstruirana matrica naziva se ***position-specific scoring matrix*** ili, kraće, **PSSM matrica**. U duhu hrvatskog jezika možemo je prevesti kao *matrica pozicijsko specifičnih težina*.

Uočimo, PSSM matricu konstruirali smo na način da smo pretpostavili nezavisnost između mjesta u nizovima, te smo za svako mjesto posebno računali vjerojatnost neovisno o drugim mjestima. To ustvari znači da vjerojatnost pojave neke aminokiseline na jednom mjestu ne ovisi o vjerojatnosti pojave te iste aminokiseline na nekom drugom mjestu. Dimenzija matrice jednaka je $6x4$, odnosno umnošku broja mjesta u motivu te broja aminokiselina (u ovom slučaju 4 jer motiv čine samo 4 osnovne aminokiseline). Iz toga je lako zaključiti da, ukoliko bismo radili profil motiva koji se sastoji od nizova duljine 6 u kojima je zastupljeno svih 20 aminokiselina, tada bi dimenzija te matrice bila $6x20$.

T. D. Schneider i R. M. Stephens su 1990. godine predstavili grafičku metodu prikazivanja motiva. Grafički prikaz sastoji se od niza likova. Broj likova jednak je broju mjesta u motivu, a svaki se pojedini lik sastoji od slova koja se nalaze na određenom mjestu. Visina tih slova u likovima proporcionalna je njihovoj frekvenciji pojavljivanja na određenim mjestima. Takav grafički prikaz nizova poznat je još i pod imenom *sequence logo*.

Grafički prikaz našeg motiva prikazan je na slici[2]:

[2] *Sequence logo* izrađen na http://weblogo.berkeley.edu/logo.cgi.



Kako bismo još bolje dobili sliku o toj grafičkoj metodi prikažimo i motiv enzima kojeg smo također analizirali na početku:

PSSM matricu prvi je, sa svojim suradnicima, 1982. godine uveo američki genetičar Gary Stormo te se je ona pokazala vrlo uspješnom za otkrivanje motiva, kao i za proučavanje familije nizova te traženje sličnosti među njima.

Kostruirajmo sada općenitu PSSM matricu za neki motiv:

Neka je $\mathcal{A}$ = [`A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V`], te neka motiv sadrži n poravnatih nizova dujine L.
Poravnate nizove koji čine motiv označimo na sljedeći način:

$$\begin{matrix} x_{1,1}\ x_{1,2}\ ...\ x_{1,L} \\ x_{2,1}\ x_{2,2}\ ...\ x_{2,L} \\ \vdots \\ x_{n,1}\ x_{n,2}\ ...\ x_{n,L} \end{matrix}$$

gdje je $x_{i,j} \in \mathcal{A}, i = 1, 2, ..., n, j = 1, 2, ..., L$.

Position-specific scoring matrix ili, kraće, PSSM matricu označimo s

$$M = [p_{i,j}], i = 1, 2, ..., L, j = 1, 2, ..., 20.$$

Definirajmo funkciju

$$\delta_{a_j}(x_{i,j}) = \begin{cases} 1, & a_j = x_{i,j} \\ 0, & a_j \neq x_{i,j} \end{cases}$$

gdje je $a_j \in \mathcal{A}, j = 1, 2, ..., 20, i = 1, 2, ..., n$.

Tada elemente matrice M računamo kao

$$p_{i,j} = \frac{1}{n} \sum_{i=1}^{n} \delta_{a_j}(x_{i,j}), \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., 20. \tag{4.1}$$

4.3 *Score* poravnanja pomoću PSSM-a

Pretpostavimo da smo pronašli nekoliko varijanti nekog proteina te smo na temelju njih konstruirali profil tog motiva. Ono što bi se mogli pitati je postoji li još neke varijante tog proteina. Kako bismo to saznali, trebamo provjeriti koliko je neki određeni protein poravnan s motivom, odnosno koliko mu je sličan. U tu svrhu računamo vjerojatnost poravnanja tog proteina s motivom.

Označimo li s M profil motiva, a s X proteinski niz čije poravnanje s motivom želimo izračunati, tada vjerojatnost poravnanja niza X s motivom uz uvjet M označavamo s

$$\mathbb{P}(X|M).$$

Pokažimo na primjeru jednog niza kako bismo izračunali tu vjerojatnost.

Primjer 4.1. *Profil motiva M zadan je sljedećom matricom:*

	A	C	G	T
1	0.375	0.25	0.125	0.25
2	0	0	1	0
3	0.375	0	0	0.625
4	0.125	0.25	0.25	0.375
5	0.25	0.25	0.25	0.25
6	1	0	0	0

Pitamo se koja je vjerojatnost da je niz $X =$ AGACTA poravnan s motivom uz uvjet M.

Prisjetimo se, kod analize motiva razlikovali smo mjesta u nizu te nam je bilo bitno koja je vjerojatnost pojavljivanja određene aminokiseline na određenom mjestu. Iz tog razloga, prethodno postavljeno pitanje možemo preoblikovati u pitanja: koja je vjerojatnost da se aminokiselina A *iz niza X pojavi na prvom mjestu u motivu, koja je vjerojatost da se aminokiselina* G *iz niza X pojavi na drugom mjestu u motivu, itd. S obzirom da vjerojatnost pojavljivanja jedne aminokiseline na određenom mjestu ne utječe na pojavljivanje te aminokiseline na drugom mjestu, iz definicije 2.11. slijedi:*

$$\begin{aligned}\mathbb{P}(X|M) &= \mathbb{P}(AGACTA|M)\\ &= \mathbb{P}(A|M_1)\,\mathbb{P}(G|M_2)\,\mathbb{P}(A|M_3)\,\mathbb{P}(C|M_4)\,\mathbb{P}(T|M_5)\,\mathbb{P}(A|M_5)\\ &= 0.375x1x0.375x0.25x0.25x1\\ &= 0.0087890625\end{aligned}$$

gdje $\mathbb{P}(q|M_i), i = 1, 2, ..., 6$ *označava vjerojatnost pojavljivanja aminokiseline* $q \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ *na i-tom mjestu u profilu motiva M.*

Uočimo kako smo ovime izračunali vjerojatnost poravnanja niza X s motivom koji je opisan matricom M. Ono što još nismo uzeli u obzir je općenito vjerojatnost pojavljivanja neke aminokiseline u proteinskom nizu. Na primjer, uzmimo da je $\mathbb{P}(\texttt{A}) = 0.5$, $\mathbb{P}(\texttt{C}) = 0.125$, $\mathbb{P}(\texttt{G}) = 0.25$ *i* $\mathbb{P}(\texttt{T}) = 0.125$. *Tada je vjerojatnost niza X jednaka*

$$\begin{aligned}\mathbb{P}(X) &= \mathbb{P}(\texttt{AGACTA}) \\ &= \mathbb{P}(\texttt{A})\ \mathbb{P}(\texttt{G})\ \mathbb{P}(\texttt{A})\ \mathbb{P}(\texttt{C})\ \mathbb{P}(\texttt{T})\ \mathbb{P}(\texttt{A}) \\ &= 0.5x0.25x0.5x0.125x0.125x0.5 \\ &= 0.0004882813.\end{aligned}$$

Podijelimo li vjerojatnost poravnanja niza X s motivom, koji je opisan s matricom M, s vjerojatnošću pojavljivanja niza X, dobit ćemo vjerojatnost poravnanja niza X sa zadanim motivom

$$\frac{\mathbb{P}(X|M)}{\mathbb{P}(X)} = \frac{0.0087890625}{0.0004882813} = 17.9999981568$$

te logaritmiramo li dobivenu vrijednost,dobit ćemo score, odnosno veličinu kojom opisujemo poravnanje niza X sa zadanim motivom.

Score poravnanja niza X i zadanog motiva opisanog s matricom M jednak je 1.255272461.

Konstruirajmo općeniti izraz za računanje *score* poravnanja nekog niza sa zadanim motivom:

Neka je $\mathcal{A}$ = [`A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V`] te neka je R neki model koji opisuje ponašanje aminokiselina.
Neka je zadani motiv opisan matricom

$$M = [p_{i,j}], i = 1, 2, ..., L, j = 1, 2, ..., 20.$$

Niz čiji *score* poravnanja s motivom želimo izračunati označimo s

$$Y = y_1\ y_2\ ...\ y_L.$$

Tada je vjerojatnost da je niz Y poravnan s motivom opisanim s matricom M jednaka:

$$\begin{aligned}\mathbb{P}(Y|M) &= \mathbb{P}(y_1y_2...y_L|M) \\ &= \mathbb{P}(y_1|M_1)\ \mathbb{P}(y_2|M_2)\ ...\ \mathbb{P}(y_L|M_L) \\ &= \prod_{i=1}^{L} \mathbb{P}(y_i|M_i)\end{aligned} \tag{4.2}$$

gdje $\mathbb{P}(y_i|M_i), i = 1, 2, ..., L$, označava vjerojatnost pojavljivanja aminokiseline $y_i \in \mathcal{A}$ na i-tom mjestu u motivu opisanom s matricom M.

Označimo s $q_i = \mathbb{P}(y_i), i = 1, 2, ..., L$. Tada je vjerojatnost niza Y jednaka:

$$\begin{aligned} \mathbb{P}(Y) &= \mathbb{P}(y_1 y_2 ... y_L) \\ &= \mathbb{P}(y_1)\ \mathbb{P}(y_2)\ ...\ \mathbb{P}(y_L) \\ &= q_1 q_2 ... q_L = \prod_{i=1}^{L} q_i \end{aligned} \tag{4.3}$$

Omjer vjerojatnosti dobivenih u 4.2 i 4.3

$$\frac{\mathbb{P}(Y|M)}{\mathbb{P}(Y)} = \frac{\prod_{i=1}^{L} \mathbb{P}(y_i|M_i)}{\prod_{i=1}^{L} q_i} = \prod_{i=1}^{L} \frac{\mathbb{P}(y_i|M_i)}{q_i}$$

naziva se *omjer šansi* da se niz Y poravna s motivom koji je opisan matricom M.

Kako bismo dobili aditivni *scoring sistem*, logaritmiramo dobiveni izraz te na taj način definiramo *score* poravnanja niza Y i motiva opisanog s matricom M:

$$S = \sum_{i=1}^{L} \log \frac{\mathbb{P}(y_i|M_i)}{q_i}. \tag{4.4}$$

4.4 Metoda klizećeg prozora. Maksimalni *score*

U realnom svijetu vrlo je mala vjerojatnost da ćemo imati neki niz čija je duljina jednaka duljini motiva s kojim ga želimo usporediti i provjeriti njegovu sličnost sa zadanim motivom. Puno je vjerojatnija situacija u kojoj ćemo tražiti podnizove u nekom proteinu koji su slični zadanom motivu. Pokažimo jedan takav primjer.

Primjer 4.2. *Profil motiva M zadan je sljedećom matricom:*

	A	C	G	T
1	0.375	0.25	0.125	0.25
2	0	0	1	0
3	0.375	0	0	0.625
4	0.125	0.25	0.25	0.375
5	0.25	0.25	0.25	0.25
6	1	0	0	0

te je zadan niz $Y =$ `CATGGCTACGTGTAAATGG`*. Trebamo pronaći podniz ovog niza koji je nasličniji zadanom motivu.*

Sličnost nizova mjerili smo score-om poravnanja. Podniz niza Y koji je najsličniji zadanom motivu je podniz koji ima najveći score poravnanja sa zadanim motivom.

Kako bismo pronašli podniz s najvećim score-om, trebamo izračunati score-ove svih podnizova te pronaći onaj koji je najveći. Jedan od načina je da nasumično biramo podnizove duljine motiva (u našem slučaju duljine 6), te računamo njihove score-ove poravnanja sa zadanim motivom. Uzmemo li u obzir da neki proteini mogu imati i po nekoliko stotina aminokiselina, zaključit ćemo da takav način pretraživanja nije prikladan. Kako bismo bili sigurni da nismo zaboravili provjeriti niti jedan podniz danog niza, treba nam sustavan način pretraživanja.

Shematski prikažimo ideju sustavnog načina traženja podnizova:

$$
\begin{array}{ll}
CATGGCTACGTGTAAATGG & S_1 \\
CATGGCTACGTGTAAATGG & S_2 \\
CATGGCTACGTGTAAATGG & S_3 \\
\cdots & \\
CATGGCTACGTGTAAATGG & S_{14}
\end{array}
$$

Duljina motiva je 6, a duljina niza Y je 19. Sa shematskog prikaza lako možemo zaključiti da je ukupan broj podnizova jednak $19 - 6 + 1 = 14$.

Konstruirajmo algoritam za traženje podniza s najvećim *score*-om poravnanja sa zadanim motivom:

Neka je zadani motiv opisan matricom

$$M = [p_{i,j}], i = 1, 2, ..., L, j = 1, 2, ..., 20.$$

Niz u kojem tražimo podniz s najvećim *score*-om poravnanja sa zadanim motivom označimo s

$$Y = y_1\ y_2\ ...\ y_N.$$

Tada je ukupni broj podnizova duljie L niza Y jednak $N - L + 1$.
Shemtski prikaz traženja najvećeg *score*-a izgleda ovako:

$$\begin{array}{ll}
y_1\ y_2\ y_3\ \ldots\ y_L & S_1 \\
\quad y_2\ y_3\ y_4\ \ldots\ y_{L+1} & S_2 \\
\quad\ y_3\ y_4\ y_5\ \ldots\ y_{L+2} & S_3 \\
\quad\quad\ \vdots & \\
\quad\quad\ y_k\ y_{k+1}\ y_{k+2}\ \ldots\ y_{L+k} & S_k \\
\quad\quad\ \vdots & \\
\quad\quad\ y_{N-L+1}\ y_{N-L+2}\ y_{N-L+3}\ \ldots\ y_N & S_{N-L+1}
\end{array}$$

Iz 4.4 slijedi da je

$$S_k = \sum_{i=1}^{L} \log \frac{\mathbb{P}(y_{k+i-1}|M_i)}{q_{k+i-1}}, \quad k \in \{1, 2, ..., N - L + 1\}. \tag{4.5}$$

Ova metoda traženja motiva u nekom proteinu naziva se ***sliding window protocol*** ili, u duhu našeg jezika, **metoda klizećeg prozora**.

Najveći *score* ili ***maksimalni score*** poravnanja niza Y i motiva opisanog s matricom M definiramo kao

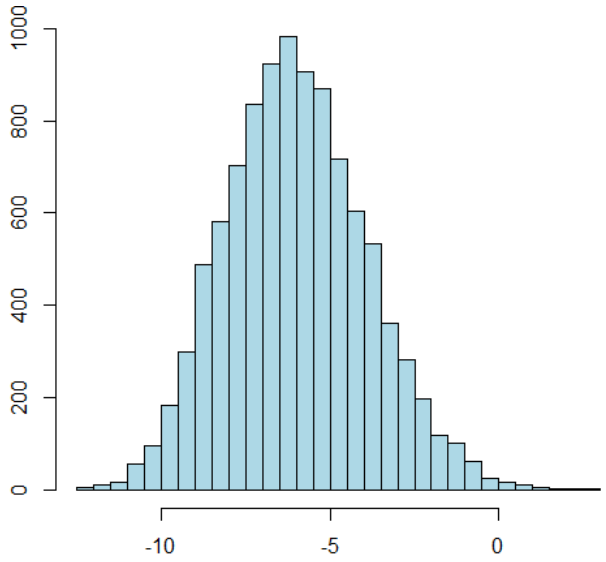
$$S = \max_{k \in \{1,2,...,N-L+1\}} S_k. \tag{4.6}$$

4.5 Distribucija *score*-va i maksimalnih *score*-ova

4.5.1 Distribucija *score*-va

Želimo analizirati *score*-ove poravnanja koji su dobiveni primjenom PSSM matrice.

Neka je zadan profil motiva opisan matricom M (PSSM matrica nalazi se u poglavlju 4.7). Simuliramo protein duljine 10000 te izračunajmo *score*-ove poravnanja kako je opisano u 4.5. Kako je duljina motiva 10, ukupan broj *score*-ova jednak je $10000 - 10 + 1 = 9991$. Distribucija tih *score*-ova prikazana je na slici:

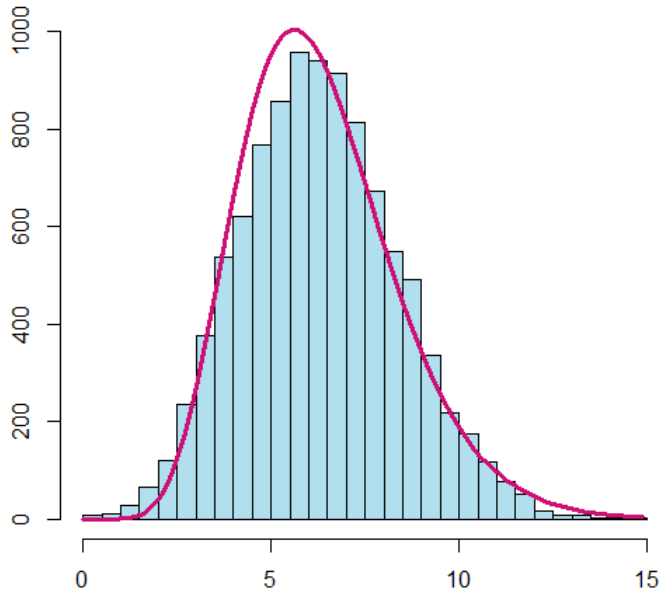


Slika 4.1: Histogram *score*-ova dobivenih pomoću *PSSM*-a

Iz dobivenog histograma na slici 4.1 možemo naslutiti da ovi *score*-ovi slijede gama distribuciju. S obzirom na definiranost gama distribuirane slučajne varijable, kako bismo dobili slučajnu varijablu koja je definirana na intervalu $\langle 0, +\infty \rangle$, dobivenim *score*-ovima pribrojimo apsolutnu vrijednost minimuma (time smo osigurali pozitivnu vrijednost svih *score*-ova).

Neka μ srednja vrijednost, a σ^2 varijanca tih *score*-ova. Dobivamo: $\mu = 6.297944$ i $\sigma^2 = 4.323431$. Iz 2.2 slijedi da je $\beta = \sigma^2/\mu$ i $\alpha = \mu/\beta$, odnosno $\beta = 9.17422$ i $\alpha = 0.6864828$.

Na sljedećoj slici 4.2 prikazani su histogram dobivenih *score*-ova i funkcija gustoće neprekidne slučajne varijable koja ima gama distribuciju s parametrima $\alpha = 0.6864828$ i $\beta = 9.17422$ u oznaci $\Gamma(\alpha, \beta)$.



Slika 4.2: Histogram *score*-ova dobivenih pomoću PSSM-a i funkcija gustoće $\Gamma(\alpha, \beta)$

4.5.2 Distribucija maksimalnih *score*-ova

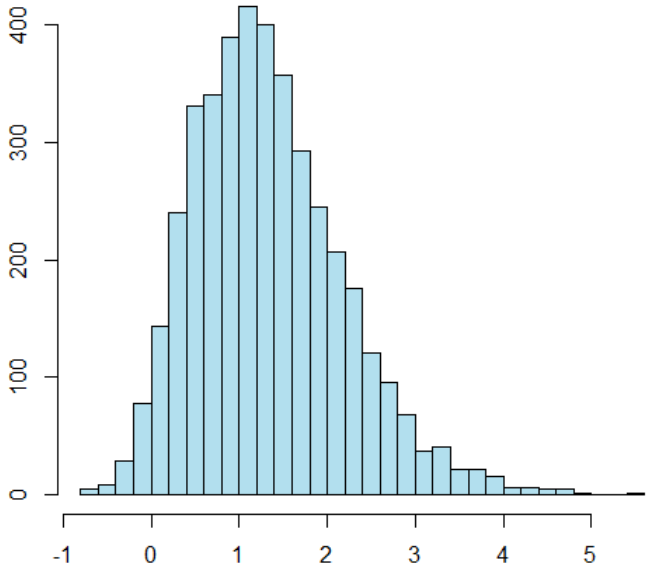
Pretpostavimo da imamo proteom koji se sastoji od nekoliko stotina proteina te želimo saznati nalazi li se u njemu neki zadani motiv. Jedan od načina je da izračunamo sve *score*-ove poravnanja sa zadanim motivom te ih usporedimo, a drugi način je da nađemo maksimalni *score* poravnanja motiva sa svakim proteinom te potom usporedimo dobivene maksimalne *score*-ove. Uočimo kako je taj postupak traženja motiva u proteomu puno jednostavniji i kraći. Iz tog razloga zanima nas kako su distribuirani maksimalni *score*-ovi.

Pretpostavimo da se neki proteom sastoji od n proteina p_n te sa S_n označimo maksimalni *score* poravnanja proteina p_n sa zadanim motivom M. Prikažimo shematski traženje zadanog motiva u proteomu:

$$
\begin{array}{ll}
p_1 & \texttt{xx...M...xxxxxxxxxxxxxx} \quad S_1 \\
p_2 & \texttt{xxxxxxx...M...xxxxxxxxx} \quad S_2 \\
p_3 & \texttt{xxxxxxxxxxx...M...xxxxx} \quad S_3 \\
 & \qquad\qquad \ldots \\
p_n & \texttt{x...M...xxxxxxxxxxxxxxx} \quad S_n
\end{array}
$$

Zanima nas kako je distribuirana varijabla $(S_n, n \in \mathbb{N})$.

Simulirajmo proteom koji sadrži 4000 proteina duljine 10000. Za svaki od proteina izračunajmo maksimalni *score* poravnanja s motivom opisanim matricom M (matrica se nalazi u poglavlju 4.7) kako je opisano u 4.6. S obzirom da se proteom sastoji od ukupno 4000 proteina, ukupan broj maksimalnih *score*-ova jednak je 4000. Uzmemo li u obzir da *score*-ovi poravnanja slijede gama distribuciju, mogli bismo pretpostaviti da će i maksimalni *score*-ovi biti gama distribuirani. Distribucija maksimalnih *score*-ova prikazana je na slici:

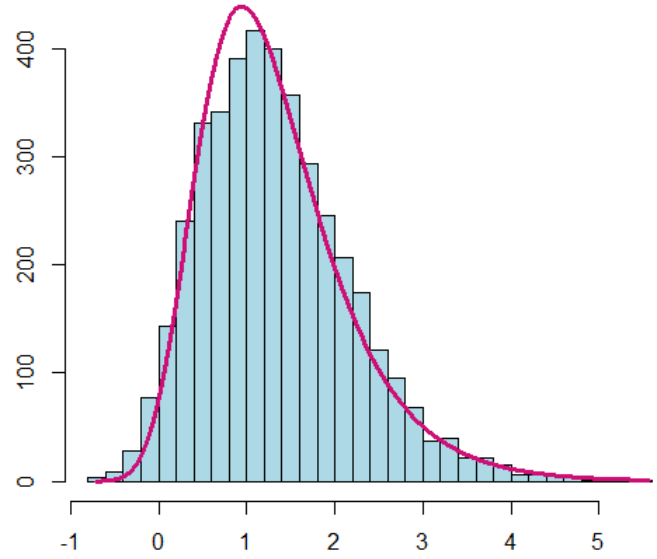


Slika 4.3: Histogram maksimalnih *score*-ova dobivenih pomoću PSSM-a

Iz histograma prikazanog na slici 4.3 lako se uoči da maksimalni *score*-ovi neće biti gama distribuirani, već će oni slijediti Gumbelovu distribuciju.

Neka je $\bar{X}$ srednja vrijednost, te S^2 uzoračka varijanca. Dobivamo $\bar{X} = 1.328995$ i $S^2 = 0.7173287$. Iz 2.3 slijedi da je $\sigma = \sqrt{\frac{6S^2}{\pi^2}}$ i $\mu = \bar{X} - \sigma\gamma$, odnosno $\sigma = 0.9478316$ i $\mu = 0.6603662$.

Na sljedećoj slici 4.5 prikazani su histogram dobivenih podataka i funkcija gustoće neprekidne slučajne varijable koja ima Gumbel distribuciju s parametrima $\sigma = 0.9478316$ i $\mu = 0.6603662$.



Slika 4.4: Histogram *score*-ova i funkcija gustoće Gumbel distribucije

Da to zaista vrijedi, potvrđuje nam *klasična teorija ekstremnih vrijednosti*.

Neka je $(X_n, n \in \mathbb{N})$ niz jednako distribuiranih slučajnih varijabli. S M_n označimo maksimum n-tog niza. Tada klasična teorija ekstremnih vrijednosti dokazuje da distribucija varijable M_n slijedi jednu od ***tri tipa distribucija ekstremnih vrijednosti***:

$$\text{Tip I:} \quad G(x) = \exp(-e^{-x}), \quad -\infty < x < \infty$$

$$\text{Tip II:} \quad G(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), \ \textit{za neko } \alpha > 0, & x > 0 \end{cases}$$

$$\text{Tip III:} \quad G(x) = \begin{cases} \exp(-(-x^{-\alpha})), \ \textit{za neko } \alpha > 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

Distribuciju ekstremnih vrijedosti Tipa I nazivamo još i *Gumbelova*[3] *distribucija*.

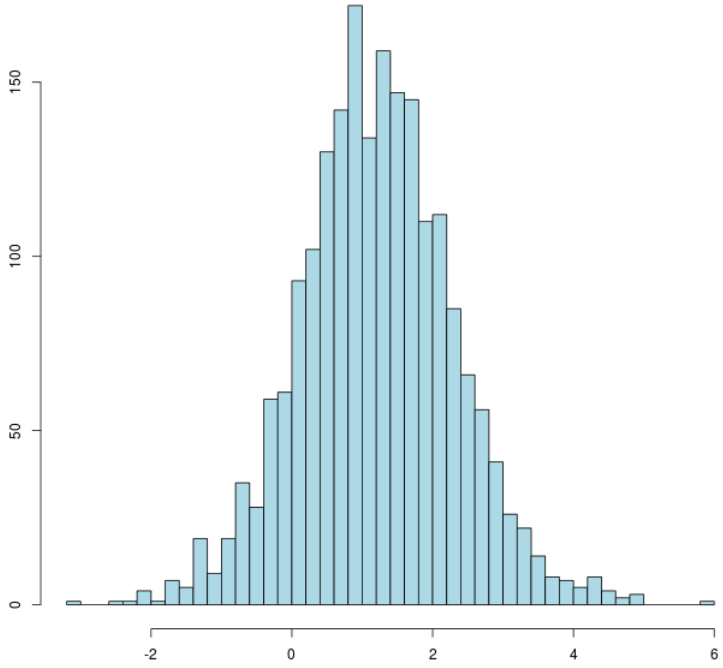
Uočimo, uvrstimo li $\mu = 0$ i $\sigma = 1$ u oćeniti oblik Gambelove distribucije, dobit ćemo standardnu Gumbelovu distribuciju Tipa I.

[3]Emil Julius Gumbel (18.7.1891.-10.9.1966.) - njemački matematičar koji se bavio distribucijom ekstremnih vrijednosti

4.6 Proteom *S. Avermitilis*

Pokažimo na primjeru stvarnog proteoma *S. Avermitilis* da maksimalni *score*-ovi slijede Gumbel distribuciju.

Neka je zadan profil motiva opisan matricom M (PSSM matrica nalazi se u poglavlju 4.7). Proteom se sastoji od 2044 proteina. Maksimalne *score*-ove poravnanja sa zadanim motivom računat ćemo kako je opisano u 4.6. Zanima nas kako su distribuirani ti maksimalni *score*-ovi.



Slika 4.5: Histogram *score*-ova

Iz dobivenog histograma uočavamo kako maksimalni *score*-ovi ovog proteoma zaista slijede Gumbel distribuciju.

4.7 Primjer PSSM matrice

1	0.0062034	0.0029790	0.0046652	0.0013492	0.0026161	0.0013318
	0.0017667	0.0032811	0.0065049	0.0132762	0.0260832	0.0021275
	0.0027851	0.6903614	0.0030683	0.0074072	0.0042897	0.0033821
	0.2113728	0.0051489				
2	0.1129740	0.0045736	0.0521998	0.0109920	0.0044169	0.0223472
	0.0104522	0.0163554	0.0064941	0.1014222	0.0725795	0.0105944
	0.0082585	0.0045076	0.0092780	0.0211381	0.0651152	0.0010448
	0.0036836	0.4615728				
3	0.0123705	0.0053575	0.0068498	0.0025712	0.0025295	0.0020967
	0.0028781	0.0069218	0.0050998	0.0822246	0.0464964	0.0042550
	0.0041289	0.6113977	0.0053198	0.0624034	0.0106797	0.0494684
	0.0608579	0.0160931				
4	0.0500473	0.0038604	0.0185282	0.0144125	0.0033589	0.0042292
	0.0108499	0.6735647	0.0020855	0.0020608	0.0040072	0.0092266
	0.0014576	0.0038226	0.0105474	0.1581025	0.0185018	0.0014420
	0.0016927	0.0082022				
5	0.0277429	0.0020932	0.1021459	0.5113153	0.0012108	0.0145841
	0.1137217	0.1400540	0.0098289	0.0037110	0.0029790	0.0180527
	0.0010965	0.0017822	0.0052141	0.0245019	0.0119220	0.0010398
	0.0020601	0.0049437				
6	0.0670035	0.0125205	0.0353388	0.0134664	0.0108544	0.0059758
	0.0107771	0.0898507	0.0041960	0.0038802	0.0044439	0.0189782
	0.0029130	0.0052780	0.0247382	0.6181021	0.0578722	0.0029702
	0.0033664	0.0074745				
7	0.0166137	0.0045800	0.0629858	0.0093604	0.0024328	0.0064908
	0.0060648	0.0076718	0.0064985	0.1193033	0.4916167	0.0108808
	0.0239445	0.0461458	0.0060514	0.0299007	0.0285520	0.0012982
	0.0461025	0.0735056				
8	0.0626144	0.0068731	0.0352657	0.0091804	0.0062689	0.0061473
	0.0076830	0.0205825	0.0294492	0.0343557	0.0881511	0.0117993
	0.0059648	0.0853645	0.0124802	0.1995867	0.0801367	0.0020387
	0.0807550	0.2153027				
9	0.0251633	0.0020478	0.0901710	0.6056042	0.0011707	0.0175496
	0.1504921	0.0254536	0.0099281	0.0039543	0.0026166	0.0184496
	0.0010907	0.0014024	0.0049585	0.0207971	0.0117286	0.0010367
	0.0019823	0.0044028				
10	0.2210329	0.0046064	0.0695206	0.0198096	0.0038109	0.0080700
	0.0363187	0.1342557	0.0064147	0.0126251	0.0102235	0.0168640
	0.0035860	0.0264458	0.0340434	0.0979153	0.2140778	0.0012926
	0.0046035	0.0744833				

Slika 4.6: Primjer PSSM matrice dimenzije $10x20$ koja je korištena u poglavlju 4.5

Poglavlje 5

Dodatak - primjena u osnovnoj i srednjoj školi

Jedna od glavnih značajki suvremene nastave matematike je koreliranost s drugim područjima znanosti i ljudske djelatnosti. Poznati matematičar Lobačevski[1] je rekao:

"Nema ni jedne matematičke grane, ma koliko ona bila apstraktna, koja se jednom ne bi mogla primijeniti na pojave stvarnog svijeta."

U suvremenom matematičkom kurikulumu postavljen je zahtjev za uspostavljanje međupredmetnih veza, tj. korelacija i integracija matematike s drugim nastavnim predmetima i odgojno-obrazovnim postignućima te zahtjev za povezivanje matematike s realnim svijetom.

Osim toga, suvremena nastava matematike podrazumijeva poticanje učenika na samostalnost, eksperimentalni i istraživački rad, primjenu tehnologije u nastavi, timski rad te teži novoj kulturi zadataka - zadacima otvorenog tipa. Također, ona treba biti orijentirana, učenicima što podrazumijeva korištenje metoda aktivne nastave. Učenici trebaju imati dominantnu ulogu pri formiranju matematičkih koncepata te uvježbavanju i ususatvljivanju nastavnog sadržaja. Učenicima treba biti omogućeno učenje takozvanom metodom otkrivanja te kreativan način ponavljanja i vježbanja naučenog. Metode aktivne nastave učenicima daju priliku za razvijanje odgovornosti za vlastiti uspjeh te napredovanje na području matematike.

Jedan od nastavnih predmeta s kojima možemo povezati matematiku je biologija. Primjenjujući nastavne sadržaje iz biologije učenicima možemo osmisliti kreativne zadatke za uvježbavanje te otkrivanje novih matematičkih koncepata.

[1] Nikolaj Ivanovič Lobačevski (1.12.1792.-24.2.1856.) - ruski matematičar koji se bavio geometrijom, osnivač neeuklidske geometrije

Cilj ovog poglavlja je dati primjere zadataka za učenike osnovne i srednje škole u kojma bismo znanje iz biologije primjenili za otkrivanje i uvježbavanje matematičkog nastavnog sadržaja te poticanje kreativnosti kod učenika.

U diplomskom radu, gledajući s biološke strane, pažnju smo usmjerili na proteine kao osnovu svih funkcija u ljudskom organizmu. Gledajući s matematičkog područja, usmjerili smo se na područje vjerojatnosti i statistike. Upravo to nam je cilj ujediniti i u ativnostima za učenike.

5.1 Osnovna škola

Prema *Nacionalnom okvirnom kurikulumu* ili, kraće, *NOK*-u, statistika i vjerojatnost su u nastavi matematike zastupljeni od prvog razreda osnovne škole pod imenom *Podatci*.

Na kraju prvog obrazovnog ciklusa (odnosno na kraju četvrtog razreda osnovne škole) od učenika se očekuje da će:

- prikupiti, razvrstati i organizirati podatke koji proizlaze iz svakodnevnog života te ih prikazati jednostavnim tablicama, piktogramima (slikovnim dijagramima) i stupčastim dijagramima
- pročitati i protumačiti podatke prikazane jednostavnim tablicama, piktogramima i stupčastim dijagramima
- prebrojiti različite ishode u jednostavnim situacijama rabeći stvarne materijale i dijagrame
- primjenjivati osnovni jezik vjerojatnosti (ishod, moguć, nemoguć, siguran, slučajan, vjerojatan, pravedna igra, nepravedna igra i slično)
- usporediti vjerojatnosti ishoda (manje vjerojatno, jednako vjerojatan, vjerojatniji).

Osim matematičkih koncepata, u *NOK*-u je, unutar matematičkih procesa, naglasak stavljen na rješavanje problema i matematičko modeliranje. Od učenika se očekuje da će postaviti i analizirati jednostavniji problem, isplanirati njegovo rješavanje odabirom odgovarajućih matematičkih pojmova i postupaka, riješiti ga te protumačiti i vrjednovati rješenje i postupak.

U nastavku slijedi primjer jedne aktivnosti previđene za učenike nižih razreda osnovne škole:

AKTIVNOST 1. **"Kreirajmo DNA!"**

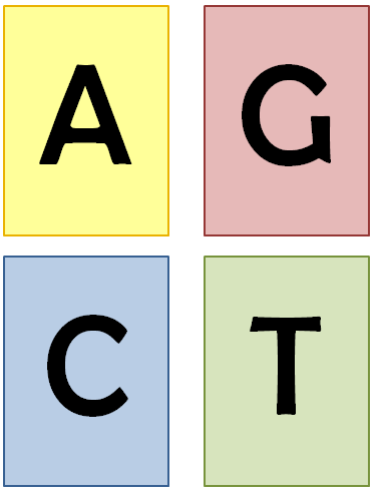
Cilj aktivnosti: učenici će, radeći u četveročlanim skupinama te konstruirajući razne DNA nizove pomoću stvarnih materijala, prebrojati različite ishode u jednostavnim situacijama te ih prikazati tablično i grafički

Oblik rada: suradničko - timski rad u četveročlanim skupinama

Nastavne metode: - heuristička nastava
- metoda eksperimenta
- metoda dijaloga

Potrebni materijal: - nastavni listić za svakog učenika (PRILOG 1)
- kartice sa slovima `A,C,G,T` (špil od $4x4$ kartice)

Tijek aktivnosti: Učenike rasporedimo u četveročlane skupine te svakoj skupini učenika podijelimo špil od 16 kartica. Svaki učenik dobije nastavni listić te ga rješava zajedno sa članovima svoje skupine. Nakon što svi učenici riješe nastavni listić, s učenicima provjerimo do kojih zaključaka su došli.



Slika 5.1: Primjer kartica sa slovima `A,C,G,T`

PRILOG 1

Kreirajmo DNA!

Je li vam netko ikad rekao “Ti si ista svoja mama!” ili “Baš me podsjećaš na svoga djeda!”?

Krivci za našu sličnost s majkama i očevima, bakama i djedovima nazivaju se **geni**.
Geni su građeni od DNA.

Zamislite jedan dugi lanac. Svaka karika tog lanca ima jedno od slova `A`,`C`,`G` ili `T`. Jedan takav lanac predstavlja **DNA**.

Bi li želio/željela znati kako izgleda tvoj DNA zbog kojih imaš plave, smeđe ili možda zelene oči?

Upusti se u pustolovinu sa svojim prijateljima iz skupine te, prateći upute, kreiraj nove DNA lance!

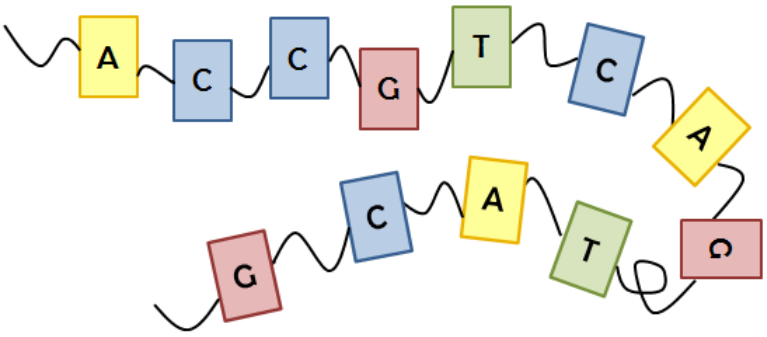
Zadatak 1. Pomoću kartica sa slovima `A`,`C`,`G` i `T` koje predstavljaju karike lanca DNA, kreirajte DNA koja ima samo dvije *različite* karike. Na koliko načina ste mogli napraviti takav DNA? Ispišite sve načine!
Redoslijed slova je bitan!

Zadatak 2. Pomoću kartica sa slovima `A`,`C`,`G` i `T` koje predstavljaju karike lanca DNA, kreirajte DNA koja ima tri *različite* karike. Na koliko načina ste mogli napraviti takav DNA? Ispišite sve načine!
Redoslijed slova je bitan!

Ako karike lanca mogu biti *jednake*, koliko biste tada različitih DNA kreirali? Isti broj, manje ili više?

Zadatak 3. Kreirajte DNA koja im četiri *različite* karike? Koliko ste mogućnosti pronašli? Jeste su li to sve mogućnsti?

Zadatak 4. Lanci DNA sadrže obično više jednakih karika. Primjer jednog takvog lanca prikazan je na slici:



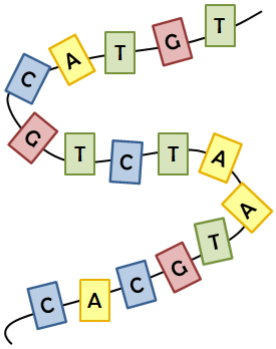
Popunite sljedeću tablicu te odgovorite na pitanja:

Naziv karike	Broj karika u DNA
A	
C	
G	
T	

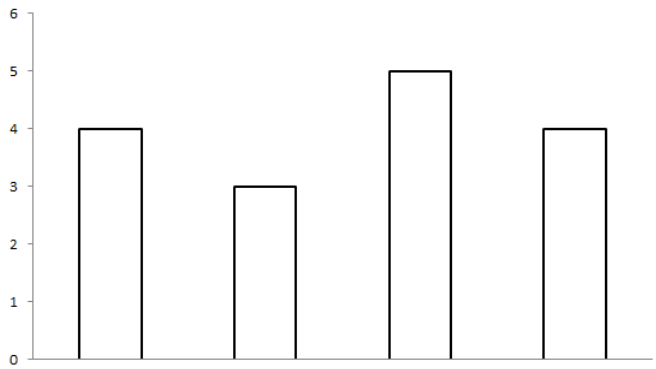
Kojih karika ima najviše, a kojih najmanje? Ima li nekih karika jednak broj?

Na temelju podataka iz tablice kreirajte novi DNA lanac!

Zadatak 5. Jedan biolog analizirao je DNA prikazanu na slici.



Kako bi sačuvao informacije o broju i vrsti karika u tom DNA lancu, nacrtao je sljedeći dijagram:



Prošlo je nekoliko godina nakon njegovog rada... Biolog je odlučio pronaći taj dijagram te se podsjetiti o broju i vrstama karika u tom lancu DNA. No, kada je pogledao dijagram nastao je problem. Zašto?

Možete li pomoći tužnom biologu popraviti njegov dijagram kako bi on ipak mogao saznati potrebne informacije?
Budite kretivni!

U nastavku školovanja, od petog do osmog razreda osnovne škole, učenici proširuju svoje znanje iz područja statistike i vjerojatnosti.

Prema *Nacionalnom okvirnom kurikulumu* na kraju osmog razreda osnovne škole, od učenika se očekuje da će :

- prikupiti, klasificirati i organizirati podatke te ih na prikladan način, pomoću računala i bez njega, prikazati sustavnom listom, tablicom, tablicom frekvencija, stupčastim i kružnim dijagramom
- pročitati, tumačiti i analizirati podatke prikazane na različite načine
- podrediti i primjeniti frekvenciju za različite podatke
- argumentirano i učinkovito odrediti broj mogućih i povoljnih ishoda u jednostavnim situacijama i izračunati vjerojatnost
- procijeniti vjerojatnost konkretnog slučajnog događaja tumačeći ju kao relativnu frekvenciju.

U nastavku slijedi prijedlog dviju aktivnosti koje su pilagođene za učenike na kraju trećeg obrazovnog ciklusa, odnosno na kraju osmog razreda:

1. Pronađite kradljivca!
2. Evolucija DNA!

AKTIVNOST 2. **"Pronađite kradljivca!"**

Cilj aktivnosti: učenici će, radeći u četveročlanim skupinama te konstruirajući razne DNA nizove i rješavajući kontekstualizirani zadatak, odrediti broj mogućih i povoljnih ishoda te izračunatu njihovu vjerojatnost

Oblik rada: suradničko - timski rad u četveročlanim skupinama

Nastavne metode: - heuristička nastava
- metoda eksperimenta
- metoda dijaloga

Potrebni materijal: - nastavni listić za svakog učenika (PRILOG 2)
- kartice sa slovima `A,C,G,T` (špil od $4x4$ kartice)

Tijek aktivnosti: Učenike rasporedimo u četveročlane skupine te svakoj skupini učenika podijelimo špil od 16 kartica. Primjer kartica prikazan je u Aktivnosti 1. Svaki učenik dobije nastavni listić te ga rješava zajedno sa članovima svoje skupine. Nakon što svi učenici riješe nastavni listić, s učenicima provjerimo do kojih zaključaka su došli.

PRILOG 2

Pronađite kradljivca!

Učenici sedmog razreda osnovne škole odlučili su štedjeti novce za izlet u NP Plitvička jezera. Svaki puta kada bi nešto uštedjeli, novce bi stavili u svoju kasicu prasicu koja se nalazi unutar njihovog razreda.

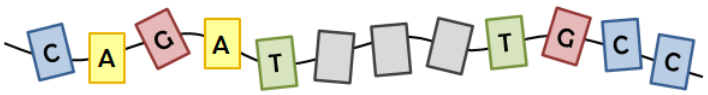
Zločesti kradljivac, vidjevši kroz prozor škole kasicu prasicu u razredu, odlučio je pod svaku cijenu doći do novaca u njoj. Jedne večeri, kada je škola bila zatvorena, kradljivac je uspio kroz prozor ući u razred. Nakon što je uzeo novce, obrisao svoje otiske ruku s prozora, misleći da ga nitko neće otkriti, sav sretan udaljio se od škole s novcima u vreći...

Je li kradljivcu bilo dosta obrisati samo svoje otiske ruku koje je ostavio na prozoru?

Gledajući kriminalističke serije jeste li ikad poželjeli biti detektiv koji će riješiti neki slučaj? Ako da, upustite se u avanturu te, zajedno sa svojim prijateljima iz skupine, pomognite detektivu pronaći kradljivca!

Osim otisaka prstiju, kradljivca možemo otkriti i po jednoj vlasi kose. Svaki naš dio tijela, pa tako i vlas kose, sadrži *dio* koji nas opisuje. Taj dio naziva se **DNA**. Zamislite jedan dugi lanac. Svaka karika tog lanca ima jedno od slova `A`,`C`,`G` ili `T`. Jedan takav lanac predstavlja DNA.

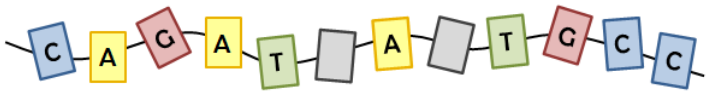
Kako bi otkrio kradljivca, detektiv treba obaviti DNA analizu. Analizirajući vlas kose, detetektiv je naišao na problem. Jedan dio DNA ne može se očitati. Prikaz DNA nalazi se na slici:



Ono što detektiv vidi iz strukture tih triju karika je da su sve one različite. Primjenjujući dobivene kartice, pronađite načine kako dopuniti DNA da dobijemo cjeloviti lanac.

1. Koliko mogućnosti ste pronašli?

2. Daljnjom analizom, detektiv je otkrio da se na srednjoj kartici nalazi slovo `A`. Koja je vjerojatnost tog ishoda?

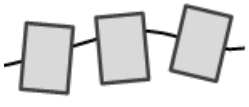


3. Na koliko načina možete popuniti ostala dva mjesta? Nađite sve mogućnosti.

4. Koja je vjerojatnost da se na preostalim karikama nalaze slova `G` i `T`, ako znamo da se na srednjoj kartici nalazi slovo `A`?

5. Vjerojatnost pojavljivanja slova `T` je 1/6, slova `G` 1/4, a slova `C` 1/3. Kako biste pomogli detektivu i smanjili broj mogućnosti za ispitivanje, odredite koji par slova ima najveću vjerojatnost pojavljivanja u nizu. Zadatak riješite grafički, pomoću stabla.

6. Dodatno, ako znate da je najvjerojatnije da će iza slova `A` doći slovo `T`, a najmanje je vjerojatno da će doći slovo `C`, koji par slova je najvjerojatniji?

Zaključak, dio DNA koji nedostaje najvjerojatnije je: 

AKTIVNOST 3. **"Evolucija DNA!"**

Cilj aktivnosti: učenici će, radeći u četveročlanim skupinama te rješavajući zadatke iz područja biologije, uvježbati određivanje frekvencije i relativne frekvencije te ih prikazati tablično i grafički

Oblik rada: suradničko - timski rad u četveročlanim skupinama

Nastavne metode: - heuristička nastava
- metoda eksperimenta
- grafička metoda
- metoda dijaloga

Potrebni materijal: - nastavni listić za svakog učenika (PRILOG 3)
- kartice sa slovima `A,C,G,T` (špil od $4x4$ kartice)

Tijek aktivnosti: Učenike rasporedimo u četveročlane skupine te svakoj skupini učenika podijelimo špil od 16 kartica. Primjer kartica prikazan je u Aktivnosti 1. Svaki učenik dobije nastavni listić te ga rješava zajedno sa članovima svoje skupine. Nakon što svi učenici riješe nastavni listić, s učenicima provjerimo do kojih zaključaka su došli.

PRILOG 3

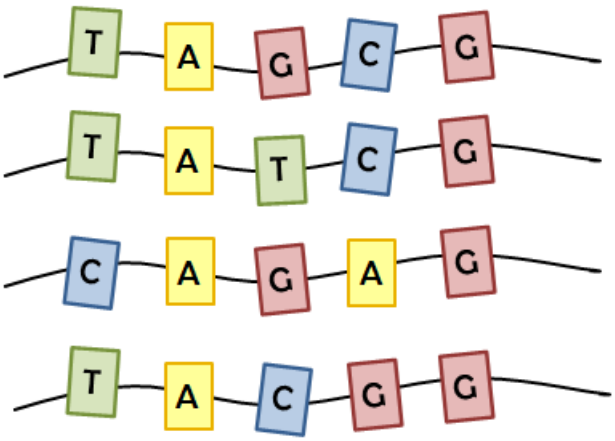
Evolucija DNA

Osnovna molekula nasljeđivanja naziva se **DNA**. DNA je odgovorna zašto Luka ima zelene oči, a Marta smeđe. Ili pak, zašto Hana ima crvenu kosu. Znate li kako je građena molekula DNA?

Zamislite jedan dugi lanac. Svaka karika tog lanca ima jedno od slova `A`,`C`,`G` ili `T`. Jedan takav lanac predstavlja DNA.

Tijekom evolucije DNA je mogla doživjeti neke promjene. Jedna od njih je zamjena jedne karike lanca drugom. Na primjer, karika sa slovom `A` zamijenjena je karikom sa slovom `G`.

Pokažimo na primjeru nekoliko varijanti jednog DNA:



Redoslijed karika u lancu je važan. Iz tog razloga razlikujemo učestalost slova na određenim mjestima u lancu.

1. Koja slova se javljaju na prvom mjestu, na drugom mjestu, na trećem mjestu, itd. u prikazanim DNA?

2. Na kojim mjestima je došlo do promjene karika u DNA lancu? Koliko puta? Postoje li mjesta na kojima nije bilo promjena?

3. Izračunajte frekvencije pojavljivanja svakog slova na karikama na pojedinom mjestu u nizovima te popunite tablicu frekvencija.

Tablica frekvencija:

Broj mjesta u nizu	A	C	G	T
1.				
2.				
3.				
4.				
5.				

Promotrite ispunjenu tablicu. Koja je vrijednost najveće frekvencije u tablici? Koja je vrijednost najmanje frekvencije u tablici? Objasnite zašto.

4. Izračunajte zbroj svih frekvencija na pojedinom mjestu. Koliko on iznosi?

Na temelju ispunjene tablice ispunite tablicu relativnih frekvencija pojavljivanja slova na pojedinom mjestu u nizu.

Tablica relativnih frekvencija:

Broj mjesta u nizu	A	C	G	T
1.				
2.				
3.				
4.				
5.				

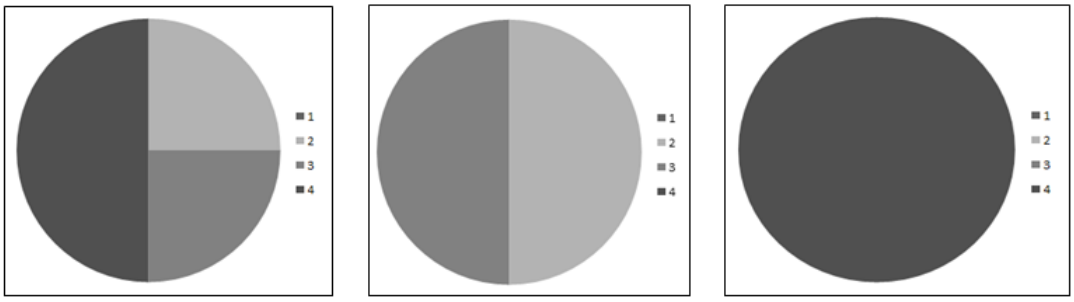
Promotrite ispunjenu tablicu. Koja je vrijednost najveće relativne frekvencije u tablici? Koja je vrijednost najmanje relativne frekvencije u tablici? Objasnite zašto.

5. Skup dobivenih DNA nastalih evolucijom nazivamo *uzorak*. Na temelju dobivenih podataka, kreirajte još neki DNA koji može biti dijelom uzorka.

Dodate li u uzorak još jedan DNA, hoće li se postojeće frekvencije i relativne frekvencije promijeniti? Istražite.

6. Pomoću stupčastog dijagrama prikažite frekvencije pojedinih slova na svakom mjestu u uzorku.

7. Prikazani su tri kružna dijagrama. Prikazuje li neki kružni dijagram frekvenciju pojavljivanja slova na određenom mjestu u uzorku? Objasnite.



8. Zadana je tablica frekvencija za neki uzorak. Koristeći kartice sa slovima kreirajte DNA lance koji čine taj uzorak. Je li uzorak koji ste kreirali jedinstven?

Broj mjesta u nizu	A	C	G	T
1.	2	0	3	0
2.	1	1	1	2
3.	0	5	0	0
4.	3	0	2	0
5.	0	0	1	4

5.2 Srednja škola

Prema nastavnom planu i programu prirodoslovno-matematičke gimnazije, učenici se u četvrtom razredu srednje škole dotiču osnovnih pojmova iz vjerojatnosti. Oni uključuju slučajne pokuse, vjerojatnosni prostor, kombinatoriku te uvjetnu vjerojatnost.

Prema *Nacionalnom okvirnom kurikulumu*, iz područja *Podatci*, od učenika se očekuje da će:

- sustavno prikupiti, klasificirati i organizirati podatke
- protumačiti složene događaje, izraziti ih pomoću skupovnih operacija te izračunati njihovu vjerojatnost.

Osim toga, unutar matematičkih procesa, područje *Rješavanje problema i matematičko modeliranje*, od učenika se očekuje da će:

- postaviti i analizirati problem, isplanirati njegovo rješavanje odabirom odgovarajućih matematičkih pojmova i postupaka, riješiti ga, te protumačiti i vrjednovati rješenje i postupak
- modelirati situacije i procese iz drugih odgojno-obrazovnih područja te svakodnevnog osobnog, profesionalnog i društvenog života.

U nastavku slijedi primjer jedne aktivnosti u kojima će učenici otkriti princip uzastopnog prebrojavanja:

AKTIVNOST 4. **"Koliko ima različitih proteina?"**

Cilj aktivnosti: učenici će, radeći u četveročlanim skupinama te rješavajući zadatke iz područja biologije, otkriti princip uzastopnog prebrojavanja

Oblik rada: suradničko - timski rad u četveročlanim skupinama

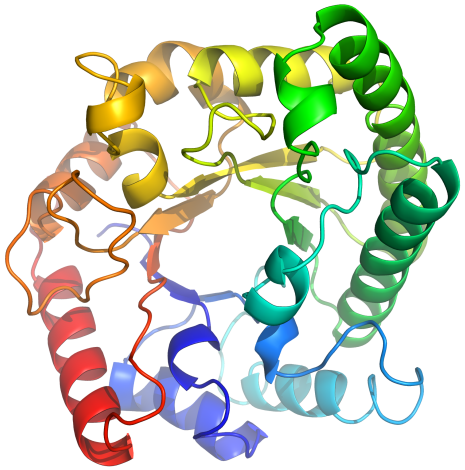
Nastavne metode: - heuristička nastava
- metoda eksperimenta
- metoda analogije i generalizacije
- metoda dijaloga

Potrebni materijal: nastavni listić za svakog učenika (PRILOG 4)

Tijek aktivnosti: Učenike rasporedimo u četveročlane skupine. Svaki učenik dobije nastavni listić te ga rješava zajedno sa članovima svoje skupine. Nakon što svi učenici riješe nastavni listić, s učenicima provjerimo do kojih zaključaka su došli.

PRILOG 4

Koliko ima različitih proteina?



Najvažnije tvari u ljudskom organizmu, uz vodu, su **proteini** koje još nazivamo i bjelančevine. Proteini su izvor tvari za izgradnju mišića, krvi, kože, kose, noktiju i unutarnjih organa, uključujući srce i mozak, te su najvažniji čimbenici u rastu i razvoju svih tjelesnih tkiva.

Proteini su molekule građene od samo 20 različitih aminokiselina koje su povezane peptidnom vezom u dugi lanac. Njihov oblik možemo zamisliti kao karike povezane u lanac, pri čemu svaka karika predstavlja jednu aminokiselinu. Promijenimo li samo jednu kariku u tom lancu, tj. neku aminokiselinu zamijenimo drugom, možemo dobiti novi protein.

Možemo se pitati kako je moguće da se s 20 različitih aminokiselina mogu načiniti proteini potrebni amebi, vrapcu ili čovjeku? Kako je moguće da su od samo 20 aminokiselina izgrađeni svi proteini živog svijeta?

Aminokiseline obično označavamo velikim tiskanim slovima

`A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V.`

Zadatak 1. Zamislite da se protein sastoji od samo dvije aminokiseline. Izaberite jednu aminokiselinu. Na koliko načina možete kreirati protein duljine 2 sa samo jednom aminokiselinom?

Zadatak 2. Izaberite dvije različite aminokiseline. Na koliko načina možete kreirati protein duljine 2 s dvije različite aminokiseline? Ispišite sve mogućnosti.

Napomena: Aminokiseline u proteinu mogu se ponavljati te je bitan redoslijed aminokiselina!

Zadatak 3. Izaberite tri različite aminokiseline. Na koliko načina možete kreirati protein duljine 2 s tri različite aminokiseline? Ispišite sve mognućnosti.

Zadatak 4. Računajući broj mogućnosti kreiranja proteina jeste li uočili neku pravilnost? Popunite sljedeću tablicu:

Duljina proteina	**Broj različitih aminokiselina koje čine protein**	**Broj kreiranih proteina**
2	1	
2	2	
2	3	
2	4	
2	5	

Broj kreiranih proteina zapišite u obliku umnoška jednakih faktora. Uočavate li sada neku pravilnost?

Koliko biste proteina duljine 2 mogli kreirati s 9 različitih aminokiselina, a koliko s 20 različitih aminokiselina?

Zadatak 5. Zamislite da se protein sastoji od samo tri aminokiseline. Izaberite jednu aminokiselinu. Na koliko načina možete kreirati protein duljine 3 sa samo jednom aminokiselinom?

Zadatak 6. Izaberite dvije različite aminokiseline. Na koliko načina možete kreirati protein duljine 3 s dvije različite aminokiseline? Ispišite sve mognućnosti.

Zadatak 7. Uočavate li neku pravilnost? Popunite sljedeću tablicu:

Duljina proteina	**Broj različitih aminokiselina koje čine protein**	**Broj kreiranih proteina**
3	1	
3	2	
3	3	
3	4	

Koliko biste proteina duljine 3 kreirali sa 7 različitih aminokiselina, a koliko s 20 različitih aminokiselina?

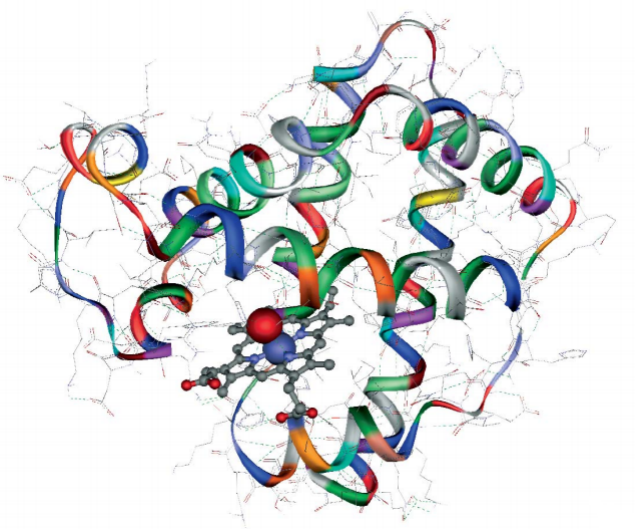
Zadatak 8. Neki proteini građeni su i od nekoliko tisuća aminokiselina. Koliko biste različitih proteina duljine 2014 kreirali s 20 različitih aminokiselina?

Mislite li još uvijek da se sa samo 20 aminokiselina ne mogu načiniti svi proteini živog svijeta?

Dvadeset različitih aminokiselina čine abecedu proteina. Isto kao što 30 različith slova čine abecedu hrvatskog jezika. Možete li sa samo 30 slova reći sve što želite?

Isto kao što se u riječima ponavljaju ista slova, tako se i u proteinima ponavljaju iste aminokiseline.

Dakle, sa samo 20 aminokiselina moguće je načiniti sve proteine živog svijeta!



Slika prikazuje strukturu mioglobina, proteina koji je zadužen za skladištenje i prijenos kisika u mišićima.

U drugom poglavlju diplomskog rada, kao primjer neprekidne slučajne varijable, naveli smo Gaussovu ili normalnu distribuciju. S Gaussovom krivuljom učenici se susreću u četvrtom razredu srednje škole. Prema *Nacionalnom okvirnom kurikulumu*, iz područja *Podatci*, od učenika se očekuje da će primijeniti normalnu razdiobu.

Osim već spomenutog matematičkog modeliranja na koje je stavljen naglasak u *NOK*-u, unutar matematičkih procesa, iz područja *Primjena tehnologije*, od učenika se očekuje da će:

- istraživati i analizirati matematičke ideje, eksperimentirati s njima te provjeravati pretpostavke pomoću džepnog računala i raznovrsnih računalnih programa, naročito programa dinamične geometrije i programa za izradbu proračunskih tablica
- razložno i učinkovito rabiti džepno računalo za računanje i tehnologiju za prikupljanje, organiziranje, prikazivanje, predstavljanje i razmjenu podataka i informacija, za rješavanje problema i modeliranje te u situacijama kojima su u središtu zanimanja matematičke ideje (radi rasterećivanja od računanja i grafičkog prikazivanja).

U nastavku se nalazi primjer aktivnosti u kojoj će učenici otkriti oblik Gaussove krivulje provodeći eksperiment te istražujući uz pomoć raznovrsnih računalnih programa:

AKTIVNOST 5. **"Gaussova krivulja"**

Cilj aktivnosti: učenici će, radeći u parovima te bacanjem simetričnog novčića, otkriti oblik Gaussove krivulje

Oblik rada: suradnički rad u parovima

Nastavne metode: - heuristička nastava
- metoda eksperimenta
- metoda crtanja
- metoda dijaloga

Potrebni materijal: - nastavni listić za svaki par učenika (PRILOG 5)
- PC, alat dinamične geometrije
- simetrični novčić za svaki par učenika

Tijek aktivnosti: Učenike rasporedimo u parove. Svaki par učenika dobije nastavni listić i simetrični novčić. Učenici u parovi ispunjavaju nastavni listić, te koriste računalne programe kako bi prikazali dobivene podatke i istražili oblik Gaussove krivulje. Nakon što svi učenici riješe nastavni listić, s učenicima provjerimo do kojih zaključaka su došli.

PRILOG 5

Gaussova krivulja

Ishodi bacanja novčića su “palo je pismo” i “pala je glava”. Kraće ih označavamo slovima P i G.

Zadatak 1. Zajednički u paru provedite jednostavan pokus bacanja simetričnog novčića. Jedan učenik baca novčić u zrak, a drugi učenik bilježi na koju stranu je novčić pao - pismo ili glava.

Jedan pokus sastoji se od deset bacanja simetričnog novčića. Provedite deset takvih pokusa. Nakon svakog izvršenog pokusa prebrojite koliko puta je pala glava u tom pokusu.

Primjer, rezultat jednog pokusa je GGPGPPGGGP. U ovom pokusa glava je pala šest puta.

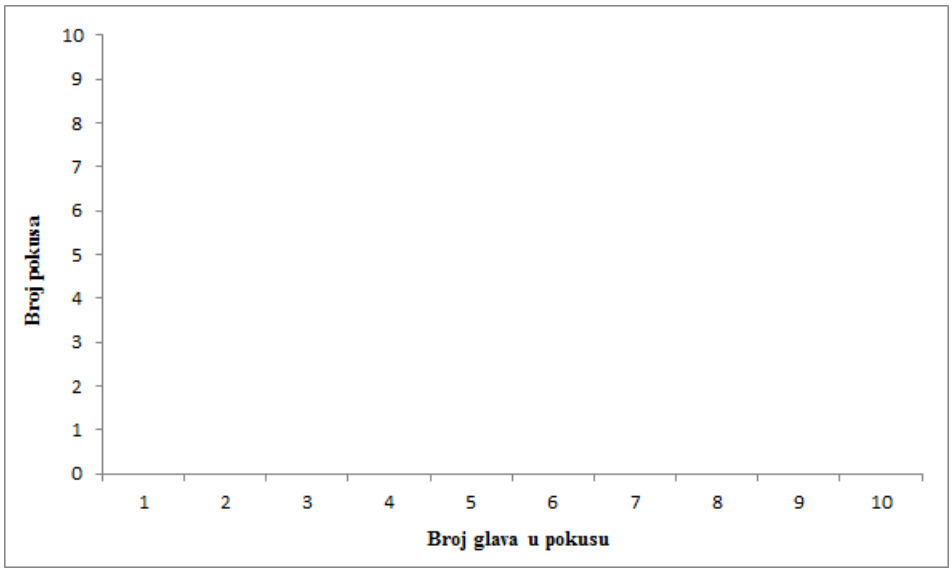
Svoje rezultate zabilježite u tablicu:

BROJ POKUSA	ISHOD POKUSA	BROJ GLAVA
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		
9.		
10.		

Zadatak 2. Koliko najviše glava je moglo pasti u jednom pokusu? A najmanje? Ispunite tablicu frekvencija broja glava u provedenim pokusima.

Broj glava u pokusu	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Broj pokusa											

Zadatak 3. Na temelju tablice frekvencija nacrtajte histogram.

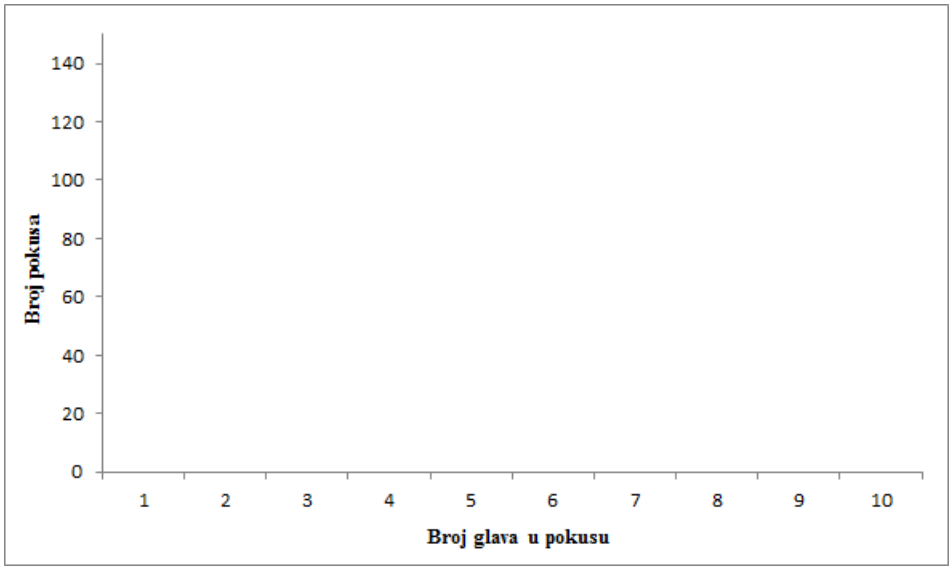


Nacrtajte krivulju koja prolazi "vrhovima" histograma. Podsjeća li vas dobivena krivulja na nešto?

Zadatak 4. Drugi parovi u razredu su provodili također isti pokus. Prikupite podatke o broju glava u pokusima od svih parova u razredu te ih unesite u sljedeću tablicu:

Ukupan broj glava u pokusu	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ukupan broj pokusa											

Na temelju prikupljenih podataka nacrtajte histogram.



Nacrtajte krivulju koja prolazi "vrhovima" histograma. Podsjeća li vas sada nacrtana krivulja na nešto?

Krivulja koju ste nacrtali naziva se još i zvonolika ili **Gaussova krivulja**. Ime je dobila po matematičaru Gaussu.

Carl Friedrich Gauss (30.4.1777.-23.2.1855.) je veliki njemački matematičar koji je poznat po svom širokom doprinosu u matematici, fizici i astronomiji. Njegovo značenje u matematici najbolje opisuje titula koju su mu dodijelili matematičari - *princeps mathematicorum*. Neki ga čak nazivaju i *Arhimedom novog doba*.

S 19 godina Gauss je uspio konstruirati pravilni 17-terokut, a uskoro nakon toga potpuno rješava problem konstrukcije pravilnih mnogokuta.

Godine 1799. dokazuje osnovni teorem algebre koji kaže da svaka algebarska jednadžba u skupu kompleksnih brojeva ima barem jedno rješenje.

Osim toga, Gauss uvodi geometrijsku predodžbu kompleksnog broja zbog čega se kompleksna ravnina danas naziva i Gaussova ravnina.

Gauss je svoj doprinos dao i teoriji brojeva objavivši djelo *Disquisitiones aritmeticae* 1801. godine. Teoriju brojeva Gauss je znao nazvati kraljicom znanosti.

Dokaz o važnosti koju Gauss pridaje matematici možemo vidjeti iz njegove poznate rečenice:

"Matematika je kraljica svih znanosti."

Krivulju, koju danas nazivamo Gaussova krivulja, Gauss je primijenio 1809. godine u djelu *Teorija gibanja nebeskih tijela*.

Gaussovom krivuljom nazivamo graf funkcije:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},\ x \in \mathbb{R}$$

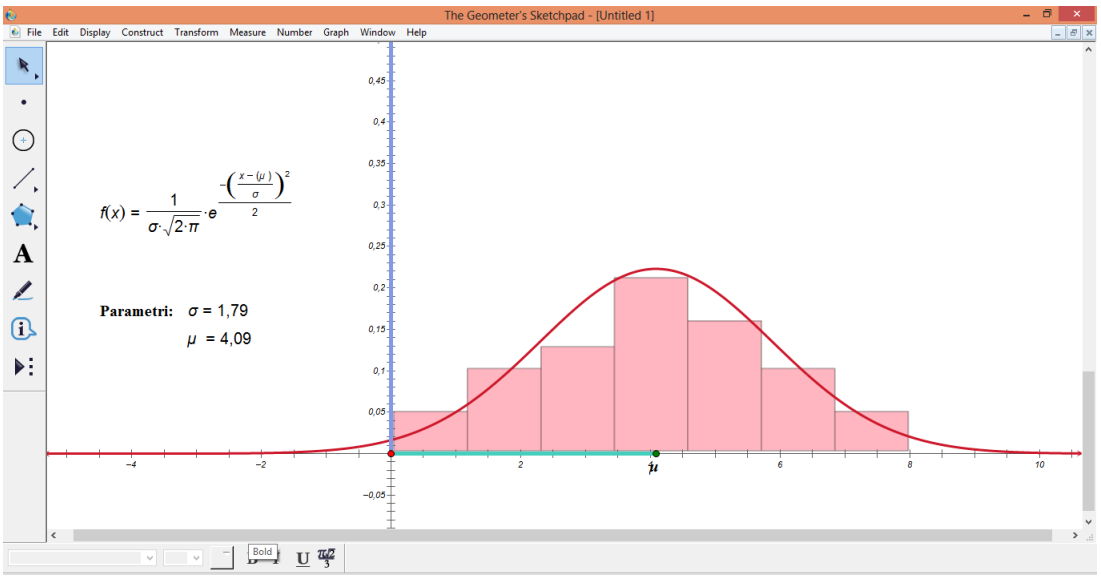
ili u općenitijem zapisu

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}},\ \mu, \sigma \in \mathbb{R}, \sigma > 0, x \in \mathbb{R}.$$

Zadatak 5. Koristeći tablicu frekvencija iz zadatka 4. napravite tablicu relativnih frekvencija. Podatke iz tablice prikažite pomoću histograma na računalu, uz pomoć nekog od programa (npr. MS Excel). Nakon toga, dobiveni histogram umetnite u neki od alata dinamične geometrije (npr. The Geometer's Sketchpad) te odredite parametre Gaussove krivulje koja najbolje *prijanja* uz dobiveni histogram.

Dobivene parametre usporedite s drugim parovima iz razreda.

U nastavku slijedi jedan primjer rješenja iz zadatka 5. Histogram je napravljen uz pomoć softvera R, a potom su parametri nađeni pomoću alata The Geometer's Sketchpad.



Bibliografija

[1] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

[2] A. M. Lesk, *Introduction to Bionformatics*, Oxford University Press, 2014.

[3] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2004.

[4] I. Slamić, *Statistička analiza DNA nizova*, Diplomski rad, PMF, Zagreb, studeni 2008.

[5] A. Relja, *Neki statistički aspekti prepoznavanja motiva*, Diplomski rad, PMF, Zagreb, srpanj 2014.

[6] S. Vrbančić, *Lokalno poravnanje i prepoznavanje motiva*, Diplomski rad, PMF, Zagreb, srpanj 2014.

[7] M. G. Leadbetter, G. Lindgren, H. Rootzén, *Extremes and related properties of random sequences and processes*, Springer-Verlag New York Heidelberg Berlin, 1983.

[8] B. Dakić, N. Elezović, *MATEMATIKA 4, udžbenik i zbirka zadataka za 4. razred gimnazije*, Element, 2007.

Sažetak

Poznata metoda traženja proteinskih motiva u nekom novom organizmu naziva se *position-specific scoring matrix* ili, kraće, PSSM metoda. U ovom diplomskom radu opisali smo PSSM algoritam te smo analizirali *score*-ove poravnanja nekog promatranog proteina i zadanog motiva koji su dobiveni primjenom PSSM metode i metodom klizećeg prozora. Došli smo do zaključka da *score*-ovi poravnanja slijede gama distribuciju, dok su maksimalni *score*-ovi Gumbel distribuirani.

Na kraju diplomskog rada pokazali smo kako uspostaviti neke moguće korelacije između biologije i matematike u osnovnoj i srednjoj školi.

Summary

In this thesis, we study PSSM - a well-known motif scanning method. We describe matrix-generation path of PSSM, as well as window-sliding algorithm that is used to obtain the optimal match. It is shown that arbitrary matches follow a gamma distribution, while the optimal scores are Gumbel distributed.

At the end, we present several interactions between teaching mathematics and biology at primary and secondary school level.

Životopis

Rođena sam 14.05.1990. godine u Koprivnici. Svoje djetinjstvo provodim u Pitomači gdje 1997. godine upisujem osnovnu školu "Petra Preradovića". Već u osnovnoškolskom obrazovanju javlja se veliki interes za matematiku. Po završetku osnovne škole, dobivam nagradu za najbolju učenicu svoje generacije. Nakon osnovne škole upisujem prirodoslovno-matematičku gimnaziju "Petra Preradovića" u Virovitici. Matematika ostaje područje mojeg najvećeg interesa. Stoga, 2009. godine upisujem preddiplomski studij Matematike, smjer: nastavnički na Prirodoslovno-matematičkom fakultetu u Zagrebu. Godine 2012. Fakultetsko vijeće i Matematički odsjek nagrađuju me za izniman uspjeh na studiju. Te iste godine stječem akademski naziv prvostupnice edukacije matematike. Nakon toga nastavljam diplomski studij Matematike, smjer: nastavnički na već spomenutom fakutetu.