Sveučilište u Zagrebu

PRIRODOSLOVNO - MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Ivo Ugrina

Hijerarhijska analiza svojstava nizova znakova metodama znanstvenog računanja i statistike

DOKTORSKI RAD

Zagreb, 2014.

University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS

Ivo Ugrina

A hierarchical analysis of character strings by statistical analysis and scientific computing

DOCTORAL THESIS

Zagreb, 2014

Sveučilište u Zagrebu

PRIRODOSLOVNO - MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Ivo Ugrina

Hijerarhijska analiza svojstava nizova znakova metodama znanstvenog računanja i statistike

DOKTORSKI RAD

Mentori:
izv.prof.dr.sc. Luka Grubišić
izv.prof.dr.sc. Bojan Basrak

Zagreb, 2014.

University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS

Ivo Ugrina

A hierarchical analysis of character strings by statistical analysis and scientific computing

DOCTORAL THESIS

Supervisors:
izv.prof.dr.sc. Luka Grubišić
izv.prof.dr.sc. Bojan Basrak

Zagreb, 2014

Zahvala

Tijekom izrade ovog doktorskog rada podršku i pomoć pružili su mi mnogi prijatelji i kolege kojima sam neizmjerno zahvalan. Dopustite da izdvojim nekoliko imena.

Prije svega zahvaljujem se svojoj obitelji, naročito *Dinki* i *Marinu*, što su me uspješno podnosili tijekom izrade disertacije usprkos mojim čestim promjena raspoloženja.

...

Sažetak

Ključne riječi: centralni granični teorem; m-zavisni nizovi; normalna distribucija; palindromi u DNA; sličnost nizova znakova; poštanske adrese; prepoznavanje adresa; geografska lokacija; stabla odlučivanja; CP dekompozicija; stršeće vrijednosti

U prvom dijelu disertacije prezentira se rezultat o distribuciji broja palindroma predodređene duljine u nizovima znakova s naglaskom na DNA nizove. Izvedeni su uvjeti pod kojima distribucija broja palindroma asimptotski teži normalnoj distribuciji. Također, izvedena je ocjena pogreške aproksimacije normalnom distribucijom te je prikazan primjer primjene na stvarnom DNA nizu.

Summary

Keywords: central limit theorem; m-dependent sequence; normal distribution; palindromes in DNA; string similarity; postal addresses; address extraction; decision trees; geographic location; CP decomposition; outliers

In the first part of this thesis a result about the distribution of the number of palindromes of a fixed length in a sequence of characters (string) is presented. Palindrome is defined as a part of the string which is equal to its complementary sequence read backwards. Complementarity is defined by some relation of characters (e.g. in DNA sequences the natural relation is C$\sim$G, A$\sim$T). The general case where the distribution of the characters is arbitrary is presented. Further, conditions in which Normal approximation can be used are derived. Special attention is given to modeling of coding and non-coding regions in DNA and to the distribution of bases in those parts of DNA. An example of an application of results to a real life sequence is also presented.

Sadržaj

Uvod

Od malih nogu društvo nas uči da je sposobnost pisanja i čitanja jako važna pa te tehnike usvajamo brzo. Ne čini nam se da u tome postoji nešto teško. Naprotiv, pisanje i čitanje nam izgledaju potpuno prirodno. Međutim, budući da smo ipak naučeni pisati i da je pisanje izrazito kompleksno, s puno pravila i još više iznimaka, često pišemo neispravno. Tako, primjerice, griješimo zbog nedostatne obrazovanosti (gramatičke i pravopisne pogreške) ili pak utjecaja okruženja u kojem živimo (dijalekti). Neispravno možemo pisati i zbog alata koje koristimo pri pisanju (pritiskanje krive tipke na tipkovnici) ili zbog utjecaja drugih kultura na razvoj alata u svakodnevnom životu (mijenjanje „č" u „c" zbog jednostavnosti zapisa u računalu). Neke od tih grešaka ne smatramo problematičnim i učestale su u jeziku dok neke druge pak smatramo velikim greškama jer mijenjaju semantiku napisane riječi.

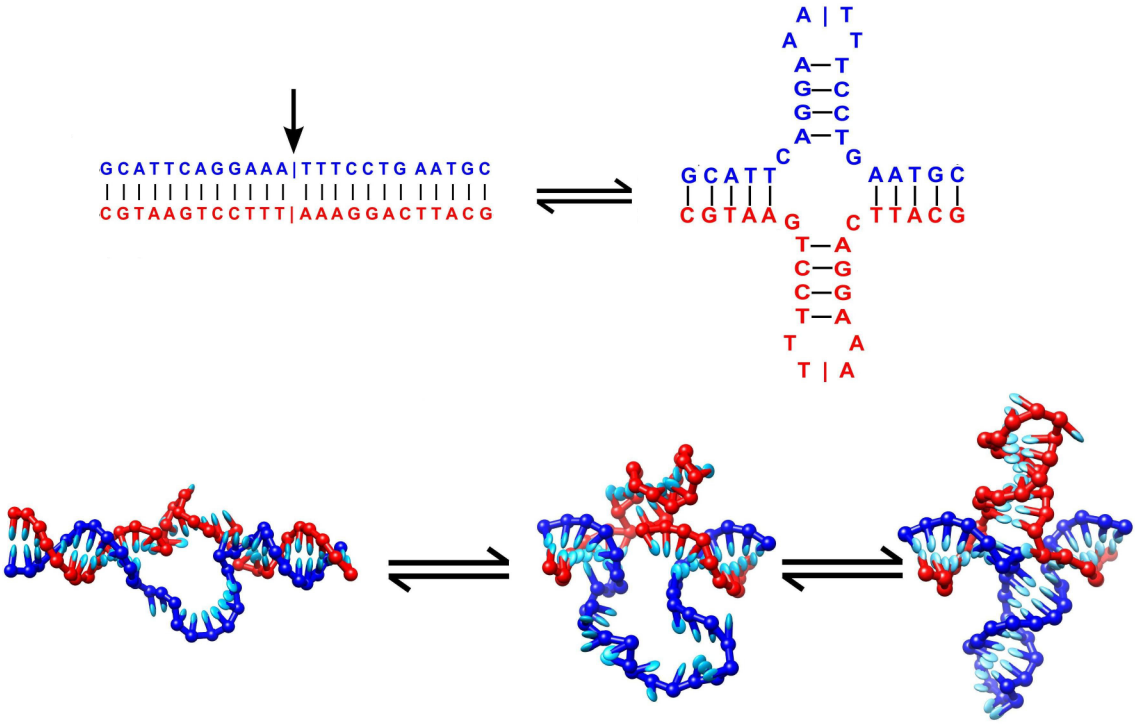
Metoda uočavanja učestalih grešaka i općenito transformacija nad riječima jako je zanimljiv aparat pri modeliranju i proučavanju prirodnih jezika. Recimo, metoda poput one koja mjeri sličnost između napisanih riječi ili općenitije nizova znakova.

No, nisu nužno zanimljive samo pogreške ili transformacije nad slovima određene riječi već i strukture koje tvori više riječi poput, primjerice, poštanskih adresa. Također, zanimljivo je i na koje se sve načine pišu poštanske adrese u kolokvijalnom pismu kao i kolika je vjerojatnost da neki niz znakova tvori poštansku adresu.

Zanimljivi mogu biti i odnosi između entiteta opisani s nekim skupom standardnih riječi poput ključnih riječi u opisima znanstvenih članaka ili pak riječi koje služe kao poveznice između WWW adresa na internetu.

Osim prirodnih jezika i pripadnih riječi, slova se danas koriste i za opisivanje drugih struktura. Primjerice, DNA nizovi se standardno prikazuju kao nizovi slova A, C, G i T. Čak se i pojam palindroma poput „Ana voli Milovana" može na odgovarajući način definirati za nizove znakova u DNA nizovima. Pogreškama u DNA nizovima možemo smatrati mutacije. Mutacije su promjene nasljedne informacije jednog organizma. Uzroci mutacija su mnogobrojni: greške pri umnožavanju genetskog materijala u procesu stanične diobe, izlaganje vanjskim čimbenicima poput radijacije i slično. Mutacije ne moraju nužno biti loše te se one u spolnim stanicama smatraju jednim od preduvjeta evolucije. Time je modeliranje pogrešaka ili transformacija nad DNA nizovima izrazito zanimljivo.

Iz rečenog nam se čini da nizovi znakova (uglavnom slova) tvore zanimljive strukture za proučavanje.



Slika 1: Stvaranja ukosnica iz palindroma u DNA nizu prikazano pomoću slovčanog i molekularnog prikaza DNA niza (izvor [16]). Strelica označava centar palindroma.

Namjera je ovog doktorskog rada prikazati svojstva nekih od tih struktura: palindroma u DNA nizovima, poštanskih adresa u hrvatskom jeziku, pogrešaka ili transformacija pri pisanju riječi te komponenata grupiranih riječima ili slovima. Točnije, namjera je matematički modelirati strukture i procese kod tih struktura.

Kao prvu zanimljivu strukturu u doktorskom radu proučavamo palindrome u DNA nizovima. Palindromi su većini poznati kao nizovi slova koji se čitaju ili izgovaraju isto od početka ili kraja. U DNA nizovima palindrome definiramo na sličan način kao i u prirodnim jezicima. Isto čitanje s oba kraja je nužno, ali uz prethodnu operaciju komplementiranja znakova. Recimo, za niz ACGT bi definiranjem komplementarnosti putem prirodnog uparivanja baza $A \sim T$ i $C \sim G$ definirali komplementaran niz sa TGCA te bi rekli da je početni niz palindrom jer se čita od početka isto kao i njemu komplementarni od kraja. Potreba za definiranjem palindroma u DNA nizovima, i proučavanje istih, može se činit čudnom. Međutim, ako primijetimo da palindromi u DNA nizu opisuju spajanje između dva dijela DNA niza onda nam motivacija postaje bliža. Naime, ako imamo palindrom u jednoj niti DNA niza onda se ta nit može spojiti sama sa sobom umjesto da se poveže s drugom niti te time tvoriti strukturu ukosnice (slika 1). Štoviše, dosadašnja su istraživanja pokazala da su palindromi u molekuli DNA nužni za funkcioniranje genoma (pogledati potpoglavlje 2.2 za reference).

Budući da su palindromi značajni za funkcioniranje genoma koristan bi bio nekakav oblik statističkog testa koji bi odavao ima li palindroma određene duljine značajno više ili

manje od očekivanog broja unutar nekog npr. roda koji se proučava. Iskaz rezultata koji omogućavaju konstrukciju takvog testa bit će prezentiran u poglavlju 2 ovog doktorskog rada.

Doprinosi doktorskog rada

- Definiran je novi model nizova znakova pomoću blokova koji se periodično ponavljaju ili zadržavaju fiksan omjer porastom duljine niza. Distribucije znakova unutar blokova su jednake, ali ne i nužno iste među blokovima. Iskazan je i dokazan teorem o asimptotskoj normalnosti broja palindroma fiksne duljine u takvim nizovima uz uvjet nezavisnosti. Dani su uvjeti na odgovarajuće procjenitelje za primjenu na nizovima kod kojih treba procjeniti parametre.

Pregled doktorskog rada po poglavljima

Prvo poglavlje naslovljeno „Pregled poznatih rezultata i tehničkih metoda" prezentira neke od najvažnijih ideja, rezultata i tehničkih metoda koje će biti potrebne u daljnjim poglavljima kao motivacija ili tehnički alat.

U drugom poglavlju naslovljenom „Distribucija broja palindroma u nizovima znakova" iskazan je i dokazan teorem o asimptotskoj normalnosti broja palindroma u nizu znakova modeliranog blokovima. Dopuštena su dva oblika blokova (periodično ponavljajući te fiksnog omjera unutar niza) s proizvoljnim distribucijama znakova unutar blokova. Nezavisnost svih znakova se pretpostavlja. Također, dani su uvjeti na procjenitelje za očekivanje i varijancu pri korištenju asimptotske normalnosti na nizove kod kojih je potrebno procjeniti parametre modela. Iznesene su i formule za očekivanje i kovarijance palindroma s proizvoljnim distribucijama. Primjerom primjene rezultata na stvaran DNA niz prikazana je razlika između n.j.d modela te modela s blokovima. Kvaliteta aproksimacije prezentirana je rezultatima simulacijske studije. Na kraju poglavlja dane su određene primjedbe i zaključak.

Oznake, kratice i slično

Pojmovi koji se prvi put spominju i od značaja su u ostatku doktorskog rada podebljani su i u kurzivu: ***tenzor***, ***procjenitelj***, ...

Skalari će se označavati malim slovima engleske ili grčke abecede: $a, b, c, \gamma, \ldots$.

Vektori će se označavati malim podebljanim slovima engleske abecede, a njihovi elementi koristit će indekse. Na primjer, v_i je i-ti element vektora $\mathbf{v} = (v_i)_{i=1}^n = (v_1, \cdots, v_n)$. U rijetkim slučajevima, zbog lakše preglednosti, koristit će se i oznaka $v[i]$ za elemente vektora.

Matrice će se označavati velikim slovima engleske abecede: $A, B, C, \ldots$. Matrični elementi označavat će se sa $M_{i,j} = M_{ij}$ ili s $M[i, j]$. Iznimno će se elementi matrice označavati drukčije (zadržavajući sintaksu i semantiku) kada elementima definiramo matricu, npr. $M = [m_{ij}]$. Za danu matricu M i-ti redak će se označavati s $M_{i,:}$ ili $M[i, :]$, a j-ti stupac s $M_{:,j}$ ili $M[:, j]$.

Tenzori će se označavati velikim podebljanim slovima engleske abecede: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots$. Elementi tenzora $\mathbf{X}$ označavat će se sa $\mathbf{X}(i_1, \cdots, i_n)$. Kada je vektor male dimenzije ponekad će se koristiti i oznaka $\mathbf{X}_{i,j,k} = \mathbf{X}_{ijk}$ za elemente tenzora.

Slučajne varijable će se označavati velikim slovima engleske abecede: $X, Y, Z, \ldots$. Budući da izrazi koji sadrže slučajne varijable uglavnom neće sadržavati i matrice dvosmislenost ne bi trebala biti problem.

Slučajni vektori će se označavati velikim podebljanim slovima engl. abecede: $\mathbf{X}, \mathbf{Y}, \ldots$. Budući da izrazi koji sadrže slučajne vektore neće sadržavati i tenzore dvosmislenost ne bi trebala biti problem. Često će se pisati $\mathbf{X} \in K$ da bi se istaknulo da je kodomena slučajnog vektora $\mathbf{X}$ jednaka K. Da slučajan vektor $\mathbf{X}$ pripada nekoj vjerojatnosnoj razdiobi η označavat će se sa $\mathbf{X} \sim \eta$. Specifično će ponekad za diskretne distribucije oznaka biti

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_K \\ p_1 & p_2 & p_3 & \cdots & p_K \end{pmatrix}$$

kada se bude htjelo naglasiti koji su mogući ishodi slučajne varijable X ($\mathrm{P}(X = a_k) = p_k$ za sve $k \leq K$).

Nizovi znakova označavat će se velikim slovima koristeći efekt malog verzala (engl. *small-caps*). Tu su mala slova zamijenjena velikim slovima manje veličine, dok su velika jednaka običnima, primjerice: Ivan Gundulić, matrica ili ACGTTTGCA. Također, u određenim će se prilikama koristiti i kurziv pri označavanju: *Ivan Gundulić*.

Nizovi u matematičkom smislu označavat će se zagradama poput $(a_k)_k$ gdje će se podrazumijevati da je $k \in \mathbb{N}$. Ukoliko bude očito po kojem se indeksu niz pomiče često će se pisati samo (a_k).

Poglavlje 1

Pregled poznatih tehničkih rezultata i metoda

U ovom će se poglavlju prezentirati neki od poznatih rezultata koji će biti od pomoći pri dokazivanju i korištenju tvrdnji vezanih uz područje interesa ovog doktorskog rada.

Autorov cilj nije u ovom poglavlju proći kroz većinu poznate literature na potrebnim područjima, već izložiti sažet i razumljiv uvod u potrebne rezultate te time omogući čitanje doktorskog rada bez nužnog posezanja za dodatnom literaturom.

Svako potpoglavlje sadrži reference na dodatnu literaturu koju zainteresirani čitatelj može konzultirati za daljnje proučavanje područja.

1.1 Konvergencija slučajnih vektora i varijabli

Za d-dimenzionalan slučajan vektor $\mathbf{X} = (X_1, \cdots, X_D)$ ***funkciju distribucije od X*** definiramo sa $F_{\mathbf{X}}(x) = \mathrm{P}(\mathbf{X} \leq \mathbf{x}) = \mathrm{P}(X_1 \leq x_1, \cdots, X_D \leq x_d)$ za sve $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$. ***Euklidsku normu*** od $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ označavamo s $|\mathbf{x}| = (x_1^2 + \cdots + x_d^2)^{1/2}$.

Definicija 1.1 Za niz slučajnih vektora $(\mathbf{X}_n)$ kažemo da konvergira slučajnom vektoru $\mathbf{X}$

(i) ***gotovo sigurno (g.s.)*** ako vrijedi

$$\mathrm{P}(\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}) = 1 \ . \tag{1.1}$$

Tada pišemo $\mathbf{X}_n \xrightarrow{g.s.} \mathbf{X}$.

(ii) ***u srednjem reda r*** ako za neki realan broj $r > 0$ te za $n \to \infty$ vrijedi

$$\mathrm{E}\left(|\mathbf{X}_n - \mathbf{X}|^r\right) \to 0 \ . \tag{1.2}$$

Tada pišemo $\mathbf{X}_n \xrightarrow{L^r} \mathbf{X}$.

(iii) ***po vjerojatnosti*** ako za svaki $\varepsilon > 0$ te za $n \to \infty$ vrijedi

$$\mathrm{P}\left(|\mathbf{X}_n - \mathbf{X}| > \varepsilon\right) \to 0 \ . \tag{1.3}$$

Tada pišemo $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$.

(iv) ***po distribuciji*** ako za sve točke x u kojima je funkcija distrubucije $F_{\mathbf{X}}$ slučajnog vektora $\mathbf{X}$ neprekidna (u oznaci $\mathbf{x} \in C(F_{\mathbf{X}})$) vrijedi

$$F_{\mathbf{X}_n}(\mathbf{x}) \to F_{\mathbf{X}}(\mathbf{x}) \tag{1.4}$$

kada $n \to \infty$. Tada pišemo $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$.

Razlog isključivanja svih točaka u kojima $F_{\mathbf{X}}$ nije neprekidna, kod konvergencije po distribuciji, možda se čini viškom. Međutim, potreban (koristan) je što prikazuje sljedeći primjer.

Primjer 1.2 Kažemo da je slučajan vektor $\mathbf{X} \in \mathbb{R}^d$ ***koncentriran*** u točki $\mathbf{c} \in \mathbb{R}^d$ ako je $\mathrm{P}(\mathbf{X} = \mathbf{c}) = 1$. Neka je $(X_n) \in \mathbb{R}$ niz slučajnih varijabli koncentriranih u točkama $1/n$ za $n = 1, 2, \ldots$ te neka je $X \in \mathbb{R}$ koncentrirana u 0. Budući da $1/n$ konvergira prema 0 kada $n \to \infty$, moglo bi se očekivati da će vrijediti $X_n \xrightarrow{D} X$.

Funkcija distribucije od X_n dana je sa $F_{X_n} = \mathbb{1}_{[1/n,\infty]}(x)$, a od X sa $F_X(x) = \mathbb{1}_{[0,\infty]}(x)$. Očito vrijedi $F_{X_n}(x) \to F_X(x)$ za $x \neq 0$, dok za $x = 0$ vrijedi $F_{X_n}(0) = 0$ i $F_X(x) = 1$. Budući da $x \notin C(F_X)$ uvjeti konvergencije po distribuciji su zadovoljeni. ∎

Sljedeća lema daje karakterizaciju konvergencije gotovo sigurno iz koje se lakše vidi razlika između konvergencije gotovo sigurno i konvergencije po vjerojatnosti.

Lema 1.3 $\mathbf{X}_n \xrightarrow{g.s.} \mathbf{X}$ ako i samo ako za svaki $\varepsilon > 0$ vrijedi

$$\mathrm{P}(|\mathbf{X}_k - \mathbf{X}| < \varepsilon,\ \forall k \geq n) \to 1 \tag{1.5}$$

kada $n \to \infty$.

Riječima razliku možemo opisati na sljedeći način: za konvergenciju po vjerojatnosti potrebno je da za svaki $\varepsilon > 0$ vjerojatnost da će se $\mathbf{X}_n$ nalaziti unutar ε odmaka od $\mathbf{X}$ teži prema 1 dok je kod konvergencije gotovo sigurno potrebno da za svaki $\varepsilon > 0$ vjerojatnost da $\mathbf{X}_k$ ostane unutar ε odmaka od $\mathbf{X}$ za svaki $k \geq n$ teži prema 1 kada n teži prema beskonačnosti.

Odnose među tipovima konvergencija slučajnih varijabli daje sljedeći teorem.

Teorem 1.4 Za niz slučajnih vektora $(\mathbf{X}_n)$ i slučajan vektor $\mathbf{X}$ vrijedi

(i) $\mathbf{X}_n \xrightarrow{g.s.} \mathbf{X} \implies \mathbf{X}_n \xrightarrow{P} \mathbf{X}$

(ii) $\mathbf{X}_n \xrightarrow{L^r} \mathbf{X} \implies \mathbf{X}_n \xrightarrow{P} \mathbf{X}$

(iii) $\mathbf{X}_n \xrightarrow{P} \mathbf{X} \implies \mathbf{X}_n \xrightarrow{D} \mathbf{X}$

Obrati ne vrijede nužno kao što pokazuju sljedeći primjeri.

Primjer 1.5 Po definiciji konvergencije po distribuciji slučajne varijable X_n ne moraju biti definirane na istom vjerojatnom prostoru kao i njihov limes X pa time ne možemo ni računati uvjet konvergencije po vjerojatnosti (1.3).

Čak i ako su X_n i X definirane na istom vjerojatnosnom prostoru konvergencija po distribuciji ne mora povlačiti konvergenciju po vjerojatnosti. Neka su $X_n \sim N(0,1)$ nezavisne i jednako distribuirane (X_n su normalne slučajne varijable s očekivanjem 0 i varijancom jednakom 1). Tada vrijedi $X_n \xrightarrow{D} X_1$ dok $X_n \not\xrightarrow{P} X_1$. ∎

Primjer 1.6 Neka je Z slučajna varijabla uniformno distribuirana na intervalu $(0,1)$, odnosno $Z \sim U([0,1])$ i neka su $X_1 = 1$, $X_2 = \mathbb{1}_{[0,1/2]}(Z)$, $X_3 = \mathbb{1}_{[1/2,1]}(Z)$, $X_4 = \mathbb{1}_{[0,1/4]}(Z)$,.... Odnosno, ako je $n = 2^k + m$, gdje je $0 \leq m \leq 2^k$ i $k \geq 0$, tada je $X_n = \mathbb{1}_{[m2^{-k},(m+1)2^{-k})}(Z)$. Očito (X_n) ne konvergira za bilo koji $Z \in [0,1)$ te time $X_n \not\xrightarrow{g.s.} X$. Međutim, $X_n \xrightarrow{L^r} 0$ za svaki $r > 0$ i $X_n \xrightarrow{P} 0$. ∎

Primjer 1.7 Neka su $Z \sim U([0,1])$ i $X_n = 2^n \mathbb{1}_{[0,1/n]}(Z)$. Tada $\mathrm{E}(|X_n|^r) = \frac{2^{nr}}{n} \to \infty$ pa vrijedi $X_n \not\xrightarrow{L^r} 0$ za svaki $r > 0$. Međutim, $X_n \xrightarrow{g.s.} 0$ ($\{\lim_{n\to\infty} X_n = 0\} = \{Z > 0\}$ i $\mathrm{P}(Z > 0) = 1$) i $X_n \xrightarrow{P} 0$ (ako je $0 < \varepsilon < 1$, $\mathrm{P}(|X_n| > \varepsilon) = P(X_n = 2^n) = 1/n \to 0$). ∎

Iako obrati u Teoremu 1.4 ne vrijede nužno, postoje dovoljni uvjeti kada obrati vrijede. Isti su dani sljedećim teoremom.

Teorem 1.8 Neka je $\mathbf{C}$ koncentriran slučajni vektor u $\mathbb{R}^d$. Tada vrijedi

(i) $\mathbf{X}_n \xrightarrow{D} \mathbf{C} \implies \mathbf{X}_n \xrightarrow{P} \mathbf{C}$.

(ii) Ako $\mathbf{X}_n \xrightarrow{g.s.} \mathbf{X}$ i $|\mathbf{X}_n|^r \leq Z$ za neki $r > 0$ i slučajnu varijablu Z sa $\mathrm{E}(Z) < \infty$ tada vrijedi $\mathbf{X}_n \xrightarrow{L^r} \mathbf{X}$.

(iii) [Scheffé (1947.)] Ako $\mathbf{X}_n \xrightarrow{g.s.} \mathbf{X}$, $\mathbf{X}_n \geq 0$ i $\mathrm{E}(\mathbf{X}_n) \to \mathrm{E}(\mathbf{X}) < \infty$ tada $\mathbf{X}_n \xrightarrow{L^1} \mathbf{X}$.

(iv) $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ ako i samo ako svaki niz prirodnih brojeva (a_k) ima podniz $(a_{p(k)})$ takav da $\mathbf{X}_{a_{p(k)}} \xrightarrow{g.s.} \mathbf{X}$ kada $k \to \infty$.

Za funkciju $g : \mathbb{R}^d \to \mathbb{R}$ kažemo da ***iščezava*** izvan kompakta ako postoji kompaktan skup $K \subset \mathbb{R}^d$ takav da vrijedi $g(\mathbf{x}) = 0$ za svaki $\mathbf{x} \notin K$.

Veza između konvergencije po distribuciji niza slučajnih vektora i konvergencije po očekivanju funkcija tih vektora dana je sljedećim teoremom.

Teorem 1.9 Sljedeći su uvjeti ekvivalentni

(i) $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$

(ii) $\mathrm{E}(g(\mathbf{X}_n)) \to \mathrm{E}(g(\mathbf{X}))$ za svaku neprekidnu funkciju g koja iščezava izvan kompakta.

(iii) $\mathrm{E}(g(\mathbf{X}_n)) \to \mathrm{E}(g(\mathbf{X}))$ za svaku ograničenu neprekidnu funkciju g.

(iv) $\mathrm{E}\,(g(\mathbf{X}_n)) \to \mathrm{E}\,(g(\mathbf{X}))$ za svaku ograničenu izmjerivu funkciju g takvu da vrijedi $\mathrm{P}(\mathbf{X} \in C(g)) = 1$.

Nekoliko korisnih pravila za rad s konvergencijom po vjerojatnosti dano je sljedećom propozicijom.

Propozicija 1.10 Pretpostavimo da za nizove slučajnih varijabli (X_n) i (Y_n) i neke konstante $a, b \in \mathbb{R}$ vrijedi $X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$ te neka je $c \in \mathbb{R}$ neka konstanta. Tada vrijedi

(i) $cX_n \xrightarrow{P} ca$,

(ii) $X_n + Y_n \xrightarrow{P} a + b$,

(iii) $X_n Y_n \xrightarrow{P} ab$,

(iv) $\dfrac{X_n}{Y_n} \xrightarrow{P} \dfrac{a}{b}$ za $b \neq 0$.

Za nizove slučajnih vektora $(\mathbf{X}_n)$ i $(\mathbf{Y}_n)$ kažemo da su ***asimptotski ekvivalentni*** ako vrijedi $(\mathbf{X}_n - \mathbf{Y}_n) \xrightarrow{P} 0$ kada $n \to \infty$.

Čest problem u teoriji velikih uzoraka (eng. *large sample theory*) je sljedeći: za dani niz slučajnih vektora $(\mathbf{X}_n)$ i dani limes po distribuciji $\mathbf{X}$ ($\mathbf{X}_n \xrightarrow{D} \mathbf{X}$) treba odrediti limes po distribuciji od $(f(\mathbf{X}_n))$ za neku funkciju f. Sljedeći teoremi prezentiraju neke od rezultata kojima se rješavaju takvi problemi.

Teorem 1.11

(i) Ako vrijedi $\mathbf{X}_n \in \mathbb{R}^d$ i $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$ te je funkcija $f : \mathbb{R}^d \to \mathbb{R}^k$ takva da $\mathrm{P}(\mathbf{X} \in C(f)) = 1$ tada slijedi $f(\mathbf{X}_n) \xrightarrow{D} f(\mathbf{X})$

(ii) Ako $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$ i $(\mathbf{X}_n - \mathbf{Y}_n) \xrightarrow{P} 0$ tada $\mathbf{Y}_n \xrightarrow{D} \mathbf{X}$.

(iii) Ako vrijedi $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$, $\mathbf{Y}_n \xrightarrow{D} \mathbf{c}$ tada

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix} \tag{1.6}$$

U prethodnom teoremu s $\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix}$ označili smo vektor iz $\mathbb{R}^{d+k}$ kojem je prvih d elemenata dano s $\mathbf{X}_n$, a zadnjih k s $\mathbf{Y}_n$, te sa $\mathbf{c}$ neki koncentrirani slučajni vektor u $\mathbb{R}^k$.

Primjer 1.12 Neka je $X \sim N(0,1)$ i $X_n \xrightarrow{D} X$. Tada za funkciju $f(x) = x^2$ po Teoremu 1.11 (dio (i)) slijedi $X_n^2 \xrightarrow{D} X^2$ budući da je f neprekidna. Budući je $X^2 \sim \chi_1^2$ kada je $X \sim N(0,1)$ dobivamo $X_n^2 \xrightarrow{D} \chi_1^2$. ∎

Primjer 1.13 Pogledajmo primjer gdje prekidnost funkcije stvara probleme. Neka je

$X_n = 1/n$ i

$$f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Tada $X_n \xrightarrow{D} 0$, ali $f(X_n) \not\xrightarrow{D} f(0)$.

■

Primjer 1.14 Dio (iii) teorema 1.11 ne može se unaprijediti pretpostavljajući da $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ te zaključivši $\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. Uzmimo $X \sim U([0,1])$ i $X_n = X$ za svaki $n \in \mathbb{N}$, $Y_n = X$ za n neparan te $Y_n = 1 - X$ za n paran. Tada $X_n \xrightarrow{D} X$ i $Y_n \xrightarrow{D} U([0,1])$, ali $\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ ne konvergira po distribuciji. ■

Izrazito važna posljedica teorema 1.11 dana je sljedećim korolarom.

Korolar 1.15 Ako vrijedi $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{Y}_n \xrightarrow{D} \mathbf{c}$ te je funkcija $f : \mathbb{R}^{d+k} \to \mathbb{R}^r$ takva da $\mathrm{P}\left(\begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix} \in C(f)\right) = 1$ tada slijedi $f(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{D} f(\mathbf{X}, \mathbf{c})$.

Primjer 1.16 Ako $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$ i $\mathbf{Y}_n \xrightarrow{D} \mathbf{c}$ iz korolara 1.15 slijedi $\mathbf{Y}_n^T \mathbf{X}_n \xrightarrow{D} \mathbf{c}^T \mathbf{X}$ budući da je skalarni produkt neprekidna funkcija. ■

Kod konvergencije po vjerojatnosti vrijede analogna pravila kao i kod teorema 1.11 osim što se dio (iii) može pojačati.

Teorem 1.17

(i) Ako vrijedi $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{X} \in \mathbb{R}^d$, $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ te za funkciju $f : \mathbb{R}^d \to \mathbb{R}^k$ vrijedi $\mathrm{P}(\mathbf{X} \in C(f)) = 1$ tada slijedi $f(\mathbf{X}_n) \xrightarrow{P} f(\mathbf{X})$.

(ii) Ako $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ i $(\mathbf{X}_n - \mathbf{Y}_n) \xrightarrow{P} 0$ tada $\mathbf{Y}_n \xrightarrow{P} \mathbf{X}$.

(iii) Ako vrijedi $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$ tada

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \tag{1.7}$$

Lako se može pokazati da vrijedi i analogon teoremu 1.17 za konvergenciju gotovo sigurno.

Jedna od osnovnih klasa teorema teorije vjerojatnosti jesu jaki i slabi zakoni velikih brojeva. Ovdje izdvajamo dva najpoznatija.

Teorem 1.18 Neka su $\mathbf{X}$, $\mathbf{X}_1$, $\mathbf{X}_2$, ... nezavisni jednako distribuirani slučajni vektori te neka je $\overline{\mathbf{X}_n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$. Tada vrijedi

$$\textbf{\textit{jaki zakon velikih brojeva:}} \quad \overline{\mathbf{X}_n} \xrightarrow{g.s.} \mathrm{E}(\mathbf{X}) \Leftrightarrow \mathrm{E}(|\mathbf{X}|) < \infty \tag{1.8}$$

$$\textbf{\textit{slabi zakon velikih brojeva:}} \quad \mathrm{E}(|\mathbf{X}|) < \infty \implies \overline{\mathbf{X}_n} \xrightarrow{P} \mathrm{E}(\mathbf{X}) \tag{1.9}$$

Vrijednost $\overline{\mathbf{X}_n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ često nazivamo ***uzoračkim očekivanjem***.

Opširniji pregled poznatih rezultata, kao i dokazi odgovarajućih tvrdnji, mogu se naći u Ferguson [4], Sarapa [25], Shiryaev i Wilson [27] ili Gut [8].

1.2 m-zavisni nizovi slučajnih varijabli

Definicija 1.19 Za niz slučajnih varijabli (Y_n) kažemo da je ***m-zavisan*** ako su za svaki $s \in \mathbb{N}$ skupovi slučajnih varijabli $\{Y_1, \cdots, Y_s\}$ i $\{Y_{m+s+1}, Y_{m+s+2}, \cdots\}$ nezavisni.

Primjetimo da je m-zavisnost ekvivalentna nezavisnosti niza slučajnih varijabli za $m = 0$.

Definicija 1.20 Za niz slučajnih varijabli Y_n kažemo da je ***stacionaran*** ako za sve $s, t \in \mathbb{N}$ distribucija slučajnog vektora $(Y_t, \cdots, Y_{t+s})$ ne ovisi o t.

Drugim riječima, niz je stacionaran ako distribucija bilo kojih s sljedećih opažanja ne ovisi o vremenu početka promatranja.

Pretpostavimo da je (Y_i) stacionaran niz m-zavisnih slučajnih varijabli te sa $\mu = \mathrm{E}(Y_1)$ označimo očekivanje od Y_1, sa $\sigma_{00} = \mathrm{Var}(Y_1)$ varijancu od Y_1 i sa $\sigma_{0i} = \mathrm{Cov}(Y_t, Y_{t+i})$ kovarijancu između Y_t i Y_{t+i}. Vrijednosti su dobro definirane i neovisne o t zbog stacionarnosti niza. Također, iz m-zavisnosti za $i > m$ slijedi $\sigma_{0i} = 0$. Očekivanja i varijance slučajnih varijabli $S_n = \sum_{i=1}^{n} Y_i$ dane su formulama

$$\mathrm{E}(S_n) = n\mu \ ,$$

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(Y_i, Y_j) = n\sigma_{00} + 2(n-1)\sigma_{01} + \cdots 2(n-m)\sigma_{0m} \ .$$

Za $n \to \infty$ vrijedi $\mathrm{Var}(S_n)/n \to \sigma^2$ gdje je $\sigma^2 = \sigma_{00} + 2\sigma_{01} + \cdots + 2\sigma_{0m}$ pa je po sljedećem teoremu distribucija od S_n aproksimativno normalna s očekivanjem μ i varijancom σ^2.

Teorem 1.21 Neka je Y_n stacionaran m-zavisan niz slučajnih varijabli s konačnim varijancama te $S_n = \sum_{i=1}^{n} Y_i$. Tada vrijedi

$$\frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{D} N(0,1), \tag{1.10}$$

odnosno

$$\frac{S_n - \mu n}{\sqrt{n}} \xrightarrow{D} N(0, \sigma^2). \tag{1.11}$$

U slučaju kada niz nije stacionaran asimptotska distribucija je također normalna uz određene uvjete. Za iskaz te tvrdnje potrebna nam je sljedeća definicija.

Definicija 1.22 Za slučajnu varijablu X definiramo slučajnu varijablu X^α kao

$$X^\alpha = \begin{cases} X, & \text{za } |X| \leq \alpha \\ 0, & \text{inače} \end{cases}$$

Za X^α kažemo da je ***odrezana*** u α.

Sljedeći nam teorem govori o konvergenciji po distribuciji trokutastog niza m-zavisnih slučajnih varijabli. Dokaz se može naći u Orey [19](Teorem 1).

Teorem 1.23 (Orey) Neka je $(p(n))$ neopadajući niz pozitivnih prirodnih brojeva koji teži prema beskonačnosti, m pozitivan prirodan broj te τ pozitivan realan broj. Ako vrijede sljedeći uvjeti:

(1) $\{X_{nv}\}$, $v = 1, \ldots, p(n)$, $n = 1, \ldots$ je trokutasti niz m-zavisnih slučajnih varijabli,

(2) $\sum_{v=1}^{p(n)} \mathrm{E}\,(X_{nv}^\tau) \to \alpha$ kada $n \to \infty$ (X_{nv}^τ su slučajne varijable X_{nv} odrezane u τ),

(3) $\sum_{v=1}^{p(n)} \sum_{w=1}^{p(n)} \mathrm{Cov}\,(X_{nv}^\tau, X_{nw}^\tau) \to \sigma^2$ kada $n \to \infty$,

(4) $\sum_{v=1}^{p(n)} \mathrm{P}(|X_{nv}| > \epsilon) \to 0$ za svaki $\epsilon > 0$,

(5) $\sum_{v=1}^{p(n)} \mathrm{Var}\,(X_{nv}^\tau) = O(1)$ kada $n \to \infty$,

tada za $n \to \infty$ vrijedi

$$\sum_{v=1}^{p(n)} X_{nv} \xrightarrow{D} N(\alpha, \sigma^2) \ .$$

Opširniji pregled poznatih rezultata, kao i dokazi odgovarajućih tvrdnji, mogu se naći u Ferguson [4].

1.3 Osnovni pojmovi matematičke statistike

Neka je Ω neprazan skup, $\mathcal{F}$ σ-algebra nad skupom Ω te $\mathcal{P}$ familija dopuštenih vjerojatnosnih razdioba nad $(\Omega, \mathcal{F})$ indeksirana nekim parametrom θ:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

gdje je $\Theta \subset \mathbb{R}^k$, $\Theta \neq \emptyset$ skup svih dopuštenih vrijednosti parametra θ neke vjerojatnosne razdiobe. Uređenu trojku $(\Omega, \mathcal{F}, \mathcal{P})$ zovemo ***statistički eksperiment*** ili ***statistički model***.

Definicija 1.24 Slučajni vektor $\mathbf{X} = (X_1, \ldots, X_n)$ čije su komponente nezavisne slučajne varijable sa zajedničkom vjerojatnosnom razdiobom, tj. za čiju funkciju distribucije vrijedi

$$\mathrm{P}_\theta(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathrm{P}_\theta(X_1 \leq x_1) \cdots \mathrm{P}_\theta(X_n \leq x_n) \ , \ \forall \theta \in \Theta, \ \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

nazivamo ***slučajnim uzorkom*** veličine n.

Definicija 1.25 ***Procjenitelj*** nepoznatog parametra θ je slučajna varijabla $\hat{T}$, definirana kao funkcija slučajnog uzorka $\mathbf{X}$. Piše se

$$\hat{T} = h(\mathbf{X}) = h(X_1, \ldots, X_n) \ ,$$

gdje je h realna funkcija n realnih varijabli.

Slučajnu varijablu kao funkciju slučajnog uzorka uobičajeno je u matematičkoj statistici zvati ***statistikom*** pa se može reći da je procjenitelj statistika.

Definicija 1.26 Procjenitelj koji zadovoljava $\mathrm{E}_\theta\left(\hat{T}\right) = \theta$ zovemo ***nepristrani procjenitelj***.

Budući da veličina uzorka u primjenama teorije statističkog zaključivanja može varirati korisno je za procjenitelj uvesti i oznaku

$$\hat{T}_n = h(X_1, \ldots, X_n), \ n \in \mathbb{N} \ ,$$

kako bi se istakla i ovisnost o veličini n. Prijašnjom je relacijom ustvari definiran beskonačan niz slučajnih varijabli $(\hat{T}_n)_{n \in \mathbb{N}}$.

Definicija 1.27 Neka je $\{P_\theta : \theta \in \Theta\}$ familija distribucija te $\mathbf{X} = \{X_1, X_2, \ldots\}$ beskonačan slučajan uzorak. Za niz procjenitelja (T_n) parametra $g(\theta)$ kažemo da je ***konzistentan*** za parametar $g(\theta)$ ako vrijedi

$$\hat{T}_n \xrightarrow{P_\theta} g(\theta), \ \forall \theta \in \Theta \ .$$

Opširniji uvod u matematičku statistiku može se naći u Pauše [20] ili Schervish [26].

Poglavlje 2

Distribucija broja palindroma u nizovima znakova

2.1 Uvod

Riječ palindrom uvriježena je među različitim nacijama kao izraz za niz znakova koji se čitaju isto slijeva nadesno i zdesna nalijevo. Prema hrvatskom jezičnom portalu ([21]) definicija bi bila „igra riječi u kojoj se čitanjem jedne riječi ili čitave rečenice obrnutim redom dobiva isto značenje kao i pravilnim čitanjem". Etimologija same riječi dolazi iz grčkog *palindromos*: koji trči natrag (*palin-* + *-drom*). Vjerojatno je većini poznat iz mozgalica kojima se djeca zabavljaju u osnovnoj školi poput „Ana voli Milovana".

Osnovno proširenje pojma palindroma dolazi iz područja bioinformatike gdje se palindromi definiraju kao nizovi znakova koji se čitaju slijeva isto kao i tome komplementarni niz zdesna. Komplementarni niz se dobiva zamjenom slova klasičnim povezivanjem adenina, citozina, guamina i timina. Odnosno, A se mijenja sa T (i obratno) te C sa G (i obratno). Primjer palindroma u DNA nizu jest ACTAGT jer je njegov komplementarni niz TGATCA. Lako je uočiti da su jedini mogući palindromi u DNA nizovima oni parne duljine.

Korisnost proučavanja palindroma u DNA nizovima slijedi iz dosadašnjih istraživanja koja ukazuju na zaključak da su palindromi u molekuli DNA nužni za funkcioniranje genoma jer se često nalaze u bis-djelujućim regijama, ali istovremeno predstavljaju ozbiljnu opasnost za genetičku stabilnost (Lobachev i dr. [11]), pri čemu genetička nestabilnost palindroma raste s njihovom duljinom (Nag i Kurst [18]). Palindromi se nalaze i kao dijelovi regulacijskih sekvenci, poput operatora (Sinden i dr. [28]), promotora (Thukral i dr. [30]), terminatora (Gomes i dr. [7]) i ishodišta replikacije (Chew i dr. [2]) te stoga čine nužan dio genetičkog materijala. Također, zbog definirane simetrije palindromi u DNA nizovima tvore sekundarne strukture poput strukture ukosnice čije formiranje tijekom replikacije može uzrokovati „proklizavanje" DNA-polimeraze te može zaustaviti replikaciju (Waldman i dr. [31]).

Višeznačnost palindroma u DNA nizovima potaknula je razna bioinformatička istra-

živanja sekvencioniranih genoma s ciljem ustanovljivanja broja, duljine i rasporeda svih palindroma.

Velik dio istraživanja sastoji se od prepoznavanja sadrži li niz značajno više palindroma od očekivanog broja (Robin i dr. [22]) pa bi poznavanje distribucije broja palindroma u unaprijed definiranom vjerojatnosnom modelu DNA niza bilo od velike koristi.

Osim u bioinformatici palindromi su standardni dio istraživanja i u diskretnoj matematici (područje koje se zove kombinatorika nad riječima) gdje, na primjer, igraju važnu ulogu pri generiranju Christoffelovih riječi s primjenama u teoriji brojeva (diofantskim aproksimacijama) (Lothaire [12] i Fischler [5]).

2.2 Povezana istraživanja

Vjerojatnosna svojstva različitih oblika nizova znakova (riječi) učestali su predmet istraživanja posljednja dva desetljeća. Povećanjem računalne moći te naročito probojem osobnih računala analiza nizova znakova, u prvom redu DNA nizova, postaje moguća te time i zanimljiva.

Značaj palindroma za različite biološke nizove, kao što je prikazano u Uvodu, dovodi do potrebe za kvalitetnim alatima pri analiziranju svojstava palindroma kao i odnosa između palindroma i nizova znakova u kojima se nalaze. Budući da su palindromi ustvari vrste riječi s posebnom strukturom, rezultate o vjerojatnosnim svojstvima riječi većinom je moguće izravno primijeniti i na palindrome.

U poglavlju „Statistike nad riječima s primjenama na biološkim nizovima" knjige [13][poglavlje 6] izvode se određena svojstva riječi pa time i palindroma. Za niz znakova pretpostavlja se da je generiran homogenim Markovljevim lancem proizvoljnog reda (n.j.d. slučaj je reda 0) unoseći time zavisnost među elementima. Međutim, ista se zavisnost podrazumijeva za cijeli niz pa se time gubi ideja regija, odnosno nije moguće modelirati dijelove niza znakova kao različite regije s različitim distribucijama. Među prezentiranim rezultatima bitno je spomenuti da su izneseni rezultati o asimptotskoj distribuciji broja palindroma kao i odgovarajućim procjeniteljima za očekivanje i varijancu. Također, dane su ocjene kvalitete aproksimacija normalnom i Poissonovom distribucijom koristeće Steinovu i Chen-Steinovu metodu. Svi su rezultati dani za homogene modele, a autori navode da do trenutka pisanja knjige nisu upoznati sa statističkim rezultatima nad heterogenim modelima.

Knjiga Robin i dr. [22] prezentacija je dijela rezultata (tehnički manje zahtjevnih) spomenutih i u knjizi [13] uz dodatak ilustracije važnosti aproksimacija broja palindroma kroz uvod u bioinformatiku.

Chew i dr. [2] promatraju problem određivanja klastera unutar DNA nizova s velikim (neočekivanim) brojem palindroma. Kao pomoć pri klasteriranju definiraju tri različite težine palindroma dajući tim određenim palindromima prednost nad drugima. U radu se

pretpostavlja homogeni model DNA niza.

Leung i dr. [9] također promatraju problem određivanja iznimnih klastera palindroma. U radu se daje dokaz nekih osnovnih svojstava palindroma u DNA nizovima te se izvodi aproksimacija pojavljivanja palindroma Poissonovim procesom za n.j.d. model DNA niza uz uvjet $p_A = p_T$ i $p_C = p_G$.

Distribucija broja palindroma po različitim regijama ljudskog genoma empirijski je određena u Lu i dr. [15].

U Chan i dr. [1] prikazana je primjena računalnih simulacija, definirani su specifični algoritmi, na određivanje distribucije broja riječi (ne nužno palindroma).

Robin i Schbath [23] daju numeričku usporedbu nekoliko aproksimacija distribucije broja riječi (međusobno i naspram egzaktne distribucije) u slučajnim nizovima znakova. Korištene su aproksimacije normalnom i složenom Poissonovom razdiobom. Pretpostavlja se homogeni Markovljev model kao model niza znakova.

Zhai i dr. [32] izvode aproksimaciju normalnom i složenom Poissonovom razdiobom za distribuciju broja riječi kod NGS (Next generation sequencing) tehnika modeliranja nizova znakova.

U diplomskom radu Gaćeša [6] razvijen je računalni program za određivanje očekivanog broja palindroma u kodirajućoj DNA. Program se zasniva na uspoređivanju broja palindroma u zadanom nizu i u računalno generiranim nizovima koji se od zadanog niza razlikuju samo po položaju sinonimnih kodona.

Autor ovog doktorskog rada, u koautorstvu s D. Špoljarićem, objavio je sažetu verziju rezultata prezentiranih u ovom poglavlju kao dio rada Špoljarić i Ugrina [29].

2.3 Matematički model

U ovom će se potpoglavlju iznijeti matematički rezultati o palindromima na proizvoljnom konačnom skupu znakova.

2.3.1 Definicija modela

Neka su slučajne varijable X_n nezavisne i distribuirane na sljedeći način:

$$X_n \sim \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_K \\ p_1^{(n)} & p_2^{(n)} & p_3^{(n)} & \cdots & p_K^{(n)} \end{pmatrix},$$

odnosno $\mathrm{P}(X_n = a_k) = p_k^{(n)}$ za sve $n \in \mathbb{N}$, $k \leq K$. Vrijednosti a_i predstavljaju znakove neke abecede, a $p_i^{(n)}$ vjerojatnosti pojavljivanja tih znakova na mjestu n u nizu. Skup svih znakova a_i označavat ćemo sa $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$.

Neka je $\widetilde{\cdot} : \mathcal{A} \to \mathcal{A}$ bijektivna funkcija takva da vrijedi $\widetilde{\widetilde{x}} = x$ za svaki $x \in \mathcal{A}$. Takvu ćemo funkciju nazivati ***funkcijom komplementarnosti***. Ukoliko budemo koristili in-

dekse kao u izrazu $\widetilde{X_{i+j}}$ često ćemo pisati samo $\widetilde{X}_{i+j}$. Najjednostavniji primjer funkcije komplementarnosti jest identiteta.

Definicija 2.1 Za niz znakova b_1, b_2,..., b_M iz $\mathcal{A}$ kažemo da tvori

(i) ***(paran) palindrom*** veličine M ako je M paran broj te vrijedi

$$b_1 = \widetilde{b}_M,\ b_2 = \widetilde{b}_{M-1}, \ldots, b_{\frac{M}{2}} = \widetilde{b}_{\frac{M}{2}+1}\ ,$$

(ii) ***neparan palindrom*** veličine M ako je M neparan broj te vrijedi

$$b_1 = \widetilde{b}_M,\ b_2 = \widetilde{b}_{M-1}, \ldots, b_{\frac{M-1}{2}} = \widetilde{b}_{\frac{M+3}{2}}, b_{\frac{M+1}{2}} = \widetilde{b}_{\frac{M+1}{2}}\ ,$$

U ostatku teksta proučavat ćemo parne palindrome te u rijetkim dijelovima i neparne palindrome.

Napomena 2.2 Iz definicije neparnog palindroma vidimo da za centralni element mora vrijediti $b_{\frac{M+1}{2}} = \widetilde{b}_{\frac{M+1}{2}}$. Stoga, broj mogućih kombinacija za neparne palindrome ovisi o bliskosti funkcije komplementarnosti identiteti $\widetilde{a}_i = a_i$ za svaki $a_i \in \mathcal{A}$. Odnosno, ovisi o broju elemenata od $a_i \in \mathcal{A}$ takvih da vrijedi $\widetilde{a}_i = a_i$. ∎

Primjer 2.3 (DNA) Poznato je da se DNA nizovi sastoje od četiri slova (A-adenin, C-citozin, G-guanin, T-timin) te da se prirodno uparuju adenin s timinom te citozin s guaninom. U uvodu smo opisali da će palindromi od značaja u DNA nizovima biti dani funkcijom komplementarnosti $\widetilde{A} = T$, $\widetilde{C} = G$, $\widetilde{G} = C$, $\widetilde{T} = A$, odnosno prirodnim uparivanjem baza. Iz napomene 2.2 slijedi da se u DNA nizovima ne mogu dobiti neparni palindromi. Primjeri parnih palindroma u DNA nizu dani su u tablici 2.1. ∎

Tablica 2.1
Primjeri palindroma u DNA nizovima

ACGT	parni palindrom duljine 4
ACCGTACGGT	parni palindrom duljine 10

Definirajmo ***blok*** znakova kao neprekinuti niz znakova duljine barem jedan. Točnije, blokom B duljine L_B smatrat ćemo bilo koji niz slučajnih varijabli oblika $(X_k, \cdots, X_{k+L_B-1})$ te ćemo zbog jednostavnosti koristiti oznaku $X_k \cdots X_{k+L_B-1}$. Promatrat će se blokovi koji su disjunktni i povezani (ne postoje elementi između blokova) te će slučajne varijable unutar svakog bloka biti jednako distribuirane.

Dva oblika povezivanja blokova su dopuštena s obzirom na povećanje niza $X_1, \ldots, X_n$ (po n):

(T1) blokovi se mijenjaju zajedno s duljinom niza n zadržavajući omjere veličina blokova unutar cijelog niza jednakima

(T2) blokovi se izmjenjuju periodično, npr. $B_1B_2B_3B_1B_2B_3\ldots$

Ideja blokova proizlazi iz potrebe da se opišu različiti dijelovi DNA nizova (npr. kodirajuće regije i nekodirajuće regije) koji mogu utjecati na broj palindroma budući da distribucije baza mogu biti različite u različitim blokovima. Na primjer, regija s visokom frekvencijom baza A i T ima veću vjerojatnost sadržavati velik broj palindroma od regije s uniformnom razdiobom baza.

U ostatku doktorskog rada podrazumijevat ćemo da su veličine blokova veće od duljine palindroma od interesa. U primjenama se takva pretpostavka podrazumijeva budući da su palindromi od interesa uvijek manji od regija koje opisuju blokovi (Lisnić i dr. [10]).

Označimo sa m fiksnu duljinu palindroma. Neka slučajna varijabla $Y_i^{(m)}$ bude indikator da palindrom duljine m završava na poziciji (indeksu) i u nizu $X_1\cdots X_n$. Odnosno, za Y_i vrijedi

$$Y_i^{(m)} = \mathbb{1}_{\{\widetilde{X}_i = X_{i-m+1}\}} \cdot \mathbb{1}_{\{\widetilde{X}_{i-1} = X_{i-m+2}\}} \cdots \mathbb{1}_{\{\widetilde{X}_{i-\frac{m}{2}+1} = X_{i-\frac{m}{2}}\}} \ . \tag{2.1}$$

Budući da je duljina palindroma jednaka m, slučajne varijable $Y_i^{(m)}$ dobro su definirane za $i \geq m$. Definirajmo sada ***broj palindroma*** duljine m kao slučajnu varijablu

$$N_n = \sum_{i=m}^{n} Y_i^{(m)} \ . \tag{2.2}$$

Ukoliko ubuduće iz konteksta bude jasno o kojoj se duljini palindroma radi, indeks m će se izostavljati iz $Y_i^{(m)}$ te će se pisati samo Y_i.

2.3.2 Osnovna svojstva palindroma

U cijelom ovom potpoglavlju duljina palindroma je fiksna i označava se sa m. Iz definicije slučajnih varijabli Y_i očito je da je $Y_m, \ldots, Y_n$ niz $(m-1)$-zavisnih slučajnih varijabli.

Lema 2.4 Za slučajne varijable Y_i vrijedi

$$\mathrm{P}(Y_i = 1) = \prod_{j=1}^{\frac{m}{2}} \left(\sum_{x\in\mathcal{A}} p_x^{(i-j+1)} p_{\widetilde{x}}^{(i-m+j)} \right)$$

Dokaz.

$$\begin{aligned}
\mathrm{P}(Y_i = 1) &= (\text{nezavisnost sl. varijabli } X_i) \\
&= \mathrm{P}(\widetilde{X}_i = X_{i-m+1}) \cdot \mathrm{P}(\widetilde{X}_{i-1} = X_{i-m+2}) \cdots \mathrm{P}(\widetilde{X}_{i-\frac{m}{2}+1} = X_{i-\frac{m}{2}}) \\
&= \left(\sum_{x\in\mathcal{A}} p_x^{(i)} p_{\widetilde{x}}^{(i-m+1)} \right) \cdot \left(\sum_{x\in\mathcal{A}} p_x^{(i-1)} p_{\widetilde{x}}^{(i-m+2)} \right) \cdots \left(\sum_{x\in\mathcal{A}} p_x^{(i-\frac{m}{2}+1)} p_{\widetilde{x}}^{(i-\frac{m}{2})} \right)
\end{aligned}$$

$$= \prod_{j=1}^{\frac{m}{2}} \left(\sum_{x \in \mathcal{A}} p_x^{(i-j+1)} p_{\widetilde{x}}^{(i-m+j)} \right)$$

□

Budući da su Y_i Bernoullijeve slučajne varijable iz prethodne leme slijede jednakosti

$$\mathrm{E}\left(|Y_i|^r\right) = \mathrm{E}\left(Y_i^r\right) = \mathrm{P}(Y_i = 1),\ r > 0 \tag{2.3}$$

$$\mathrm{Var}\left(Y_i\right) = \prod_{j=1}^{\frac{m}{2}} \left[\sum_{x \in \mathcal{A}} p_x^{(i-j+1)} p_{\widetilde{x}}^{(i-m+j)} \right] \left(1 - \prod_{j=1}^{\frac{m}{2}} \left[\sum_{x \in \mathcal{A}} p_x^{(i-j+1)} p_{\widetilde{x}}^{(i-m+j)} \right] \right) \tag{2.4}$$

čime za slučajne varijable Y_i definirane nad nezavisnim jednako distribuiranim slučajnim varijablama X_i vrijedi

$$\mathrm{P}(Y_i = 1) = \left(\sum_{x \in \mathcal{A}} p_x p_{\widetilde{x}} \right)^{\frac{m}{2}}, \tag{2.5}$$

$$\mathrm{Var}\left(Y_i\right) = \left(\sum_{x \in \mathcal{A}} p_x p_{\widetilde{x}} \right)^{\frac{m}{2}} \left(1 - \left[\sum_{x \in \mathcal{A}} p_x p_{\widetilde{x}} \right]^{\frac{m}{2}} \right) . \tag{2.6}$$

Prethodne se jednakosti mogu primijeniti na slučajne varijable X_i iz istog bloka budući da su unutar bloka X_i nezavisne i jednako distribuirane.

Za izvod zatvorene formule za kovarijancu između slučajnih varijabli Y_i i Y_j potrebne su sljedeće leme.

Lema 2.5 Za $u, v \in \mathbb{N}_0$, $u + v > 1$ vrijedi

$$\begin{aligned} &\mathrm{P}(X_{i_1} = \cdots = X_{i_u} = \widetilde{X}_{i_{u+1}} = \cdots = \widetilde{X}_{i_{u+v}}) \\ &\quad = \sum_{x \in \mathcal{A}} p_x^{(i_1)} p_x^{(i_2)} \cdots p_x^{(i_u)} p_{\widetilde{x}}^{(i_{u+1})} \cdots p_{\widetilde{x}}^{(i_{u+v})} \end{aligned} \tag{2.7}$$

Dokaz. Tvrdnja jednostavno slijedi primjenom rastava na uvjetne vjerojatnosti. □

Napomena 2.6 Za n.j.d. slučajne varijable X_i jednakost iz prethodne leme prelazi u

$$\mathrm{P}(X_{i_1} = \cdots = X_{i_p} = \widetilde{X}_{i_{u+1}} = \cdots = \widetilde{X}_{i_{u+v}}) = \sum_{x \in \mathcal{A}} \left(p_x^{(i_1)} \right)^u \left(p_{\widetilde{x}}^{(i_1)} \right)^v \tag{2.8}$$

■

Za palindrom $z_i = x_{i-m+1} \cdots x_i$ ***desni centar*** definiramo kao element $x_{i-\frac{m}{2}+1}$ te ga označavamo s $C_{z_i}^D$. ***Lijevi centar*** definiramo kao element $x_{i-\frac{m}{2}}$ te ga označavamo s $C_{z_i}^L$.

Standardne intervalne oznake koristit ćemo na sljedeći način: sa $\langle x_i, x_j \rangle$ definiramo niz elemenata od x_{i+1} do x_{j-1}; sa $\langle x_i, x_j]$ definiramo niz elemenata od x_{i+1} do x_j; sa $[x_i, x_j \rangle$ definiramo niz elemenata od x_i do x_{j-1}; sa $[x_i, x_j]$ definiramo niz elemenata od x_i do x_j.

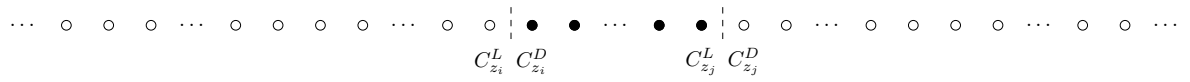
Oznake intervala koristit ćemo ponekad i kao oznake za skupove ne mareći za redoslijed znakova koje sadrže.

Nad intervalima ćemo pisati $\tilde{}$ kao oznaku za niz komplementiranih znakova originalnog niza napisan u obrnutom smjeru. Odnosno,

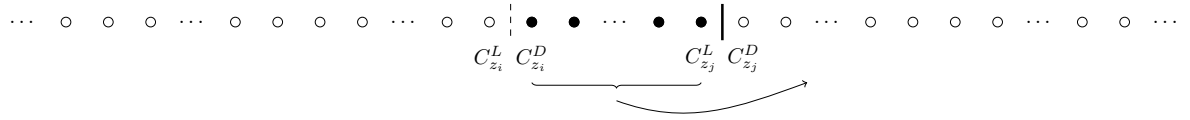
$$\widetilde{[x_i, x_j]} := \widetilde{x}_j \widetilde{x}_{j-1} \cdots \widetilde{x}_i \ .$$

Lema 2.7 Dva palindroma z_i, z_j, $j > i$ koja se sijeku ($j - i < m$) u potpunosti su određena elementima iz $\left[C^D_{z_i}, C^D_{z_j}\right\rangle$.

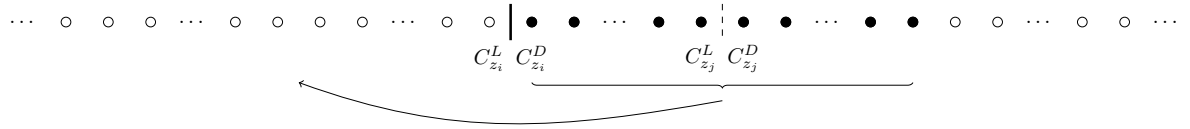
Dokaz. Ilustrirajmo dokaz opisom pravila kojim su elementi određeni kroz slike. Neka su zadani samo elementi skupa $\left[C^D_{z_i}, C^D_{z_j}\right\rangle$ (elementi ispunjeni crnom bojom).



Koristeći pravilo komplementarnosti elemenata palindroma s obzirom na centar palindroma z_j (puna linija) poznati elementi jedinstveno određuju nove elemente.



Koristeći komplementarnost elemenata palindroma z_i s obzirom na centar palindroma z_i (puna linija) poznati elementi jedinstveno određuju nove elemente.



Isti se pristup (naizmjenično korištenje komplementarnosti elemenata palindroma) sada primjenjuje sve dok se cijeli palindromi ne ispune.

□

Napomena 2.8 Iz dokaza leme 2.7 slijedi da za dva palindroma z_i i z_j koja se sijeku varijable $(X_{i-m+1}, \cdots, X_j)$ imaju sljedeću strukturu:

$$\underbrace{X_{i-m+1} \cdots X_{i-m+r}}_{\lambda \text{ ili } \widetilde{\lambda}} \cdots \lambda \quad \widetilde{\lambda} \quad \underbrace{X_{C^D_{z_i}} \cdots X_{C^D_{z_j}-1}}_{\lambda} \quad \widetilde{\lambda} \quad \lambda \cdots \quad \underbrace{X_{j-r+1} \cdots X_j}_{\lambda \text{ ili } \widetilde{\lambda}}$$

gdje λ predstavlja niz elementa $\left[C^D_{z_i}, C^D_{z_j}\right\rangle$, a $\widetilde{\lambda}$ niz $\widetilde{\left[C^D_{z_i}, C^D_{z_j}\right\rangle}$. Broj znakova u λ jest $l = j - i$, a r je ostatak pri cjelobrojnom dijeljenju $\frac{m}{2} = q \cdot l + r$. Rubni elementi sadrže samo one elemente λ ili $\widetilde{\lambda}$ koji su potrebni da bi se palindromi ispunili (duljina palindroma

ne mora nužno biti višekratnik broja elemenata u λ). Također, početak prvog palindroma i kraj drugog palindroma nalaze se u različitim nizovima (λ ili $\widetilde{\lambda}$). ∎

Neka je funkcija $S : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to \mathbb{N}^{\mathbb{N}_0}$ zadana sa

$$\begin{aligned}(S(i,j,k))(0) &= i\ ,\\ (S(i,j,k))(n) &= 2\,(j+(n-1)k) - (S(i,j,k))(n-1) - 1\ .\end{aligned} \tag{2.9}$$

Ideja u pozadini definicije funkcije S je da se na elegantan način opiše ponašanje (oblik) niza X_n iz napomene 2.8.

Definirajmo sada funkciju $\xi : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to [0,1]$ kao vjerojatnost

$$\xi(i,j,k,w) = \mathrm{P}\left(X_{S(i,j,k)(0)} = \widetilde{X}_{S(i,j,k)(1)} = X_{S(i,j,k)(2)} = \widetilde{X}_{S(i,j,k)(3)} = \ldots = \widetilde{X}_{S(i,j,k)(w)}\right)$$

za w neparan, dok ξ za w paran definiramo kao

$$\xi(i,j,k,w) = \mathrm{P}\left(X_{S(i,j,k)(0)} = \widetilde{X}_{S(i,j,k)(1)} = X_{S(i,j,k)(2)} = \widetilde{X}_{S(i,j,k)(3)} = \ldots = X_{S(i,j,k)(w)}\right)\ .$$

Vrijednosti funkcije ξ možemo računati pomoću leme 2.5.

Propozicija 2.9 Pretpostavimo da su X_n nezavisne slučajne varijable te neka je $l = j - i$. Označimo s q i r odgovarajuće koeficijente kod cjelobrojnog dijeljenja s ostatkom $m/2 = q \cdot l + r$.

a) za $l \geq m$ (Y_i i Y_j se ne sijeku) vrijedi

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \mathrm{P}(Y_i = 1)\mathrm{P}(Y_j = 1)\ ,$$

b) za $\frac{m}{2} \leq l \leq m-1$ vrijedi

$$\begin{aligned}\mathrm{P}(Y_i \cdot Y_j = 1) = &\prod_{v=i-m+1}^{i-l} \xi(v, i - \frac{m}{2} + 1, l, 2)\cdot \\ &\prod_{v=i-l+1}^{i-\frac{m}{2}} \xi(v, i - \frac{m}{2} + 1, l, 1) \prod_{v=j-l+1}^{j-\frac{m}{2}} \xi(v, j - \frac{m}{2} + 1, l, 1)\ ,\end{aligned} \tag{2.10}$$

c) za $l < \frac{m}{2}$ i $2r \leq l$ vrijedi

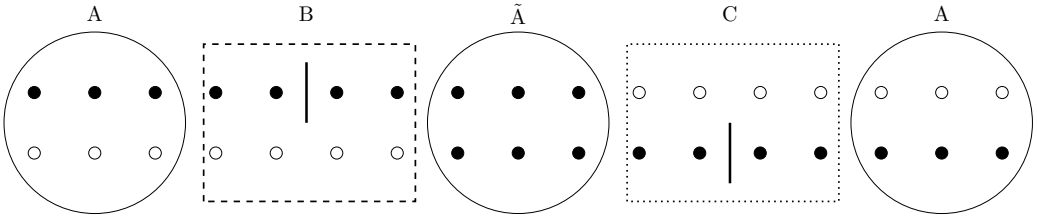
$$\begin{aligned}\mathrm{P}(Y_i \cdot Y_j = 1) = &\prod_{v=i-m+1}^{i-m+r} \xi(v, i - m + r + 1, l, 2q+1)\cdot \\ &\prod_{v=2r+1}^{l} \xi(v, i - m + r + l + 1, l, 2q) \prod_{v=l+1}^{r+l} \xi(v, i - m + r + l + 1, l, 2q+1)\ ,\end{aligned} \tag{2.11}$$

d) za $l < \frac{m}{2}$ i $2r > l$ vrijedi

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \prod_{v=i-m+1}^{i-m+l-2r} \xi(v, i-m+r+1, l, 2q+2) \cdot \prod_{v=i-m+l-2r+1}^{i-m+r} \xi(v, i-m+r+l+1, l, 2q+1) \prod_{v=l+1}^{r+l} \xi(v, i-m+r+l+1, l, 2q+1)\ . \tag{2.12}$$

Dokaz. Tvrdnja a) slijedi iz $l > m-1$ zbog nezavisnosti varijabli $\{X_{i-m+1}, \ldots, X_i\}$ i $\{X_{j-m+1}, \ldots, X_j\}$.

Zbog $\frac{m}{2} \leq l \leq m-1$ niz (X_n) ima oblik kao na slici 2.1. Iz oblika slijedi da niz A (početnih $m-l$ elemenata) mora biti komplementaran nizu $\widetilde{A}$ te isti kao i posljednjih $m-l$ elemenata. Nizovi B i C međusobno su nezavisni te nezavisni o A (odnosno $\widetilde{A}$).



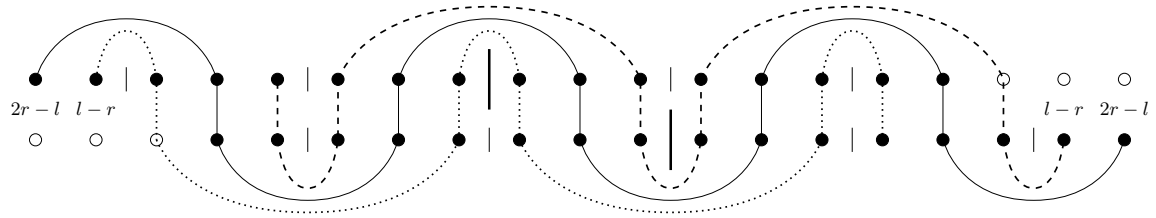
Slika 2.1: Primjer povezanosti elemenata dva presijecajuća palindroma (puni krugovi predstavljaju elemente palindroma) duljine $m = 10$ za $2r \leq l$. Podebljane linije između elemenata predstavljaju sredine palindroma.

Stoga, koristeći nezavisnost, funkciju S i lemu 2.5 dobivamo tvrdnju b).

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \prod_{t=m-1}^{l} \underbrace{\mathrm{P}\left(X_{i-t} = \widetilde{X}_{S(i-t,i-\frac{m}{2}+1,l)(1)} = X_{S(i-t,i-\frac{m}{2}+1,l)(2)}\right)}_{\text{odnos A i } \tilde{\text{A}}} \cdot \prod_{t=k}^{l-1} \left[\underbrace{\mathrm{P}\left(X_{i-t} = \widetilde{X}_{S(i-t,i-\frac{m}{2}+1,l)(1)}\right)}_{B} \underbrace{\mathrm{P}\left(X_{j-t} = \widetilde{X}_{S(j-t,j-\frac{m}{2}+1,l)(1)}\right)}_{C} \right] = \prod_{v=i-m+1}^{i-l} \xi(v, i-\frac{m}{2}+1, l, 2) \prod_{v=i-l+1}^{i-\frac{m}{2}} \xi(v, i-\frac{m}{2}+1, l, 1) \prod_{v=j-l+1}^{j-\frac{m}{2}} \xi(v, j-\frac{m}{2}+1, l, 1)$$

Slika 2.2 prikazuje slučaj $2r > l$ i $l < \frac{m}{2}$. Budući da je $2r > l$ prvih $2r-l$ elemenata nalazi se u nizu $2q+3$ puta (uz komplementarnost na odgovarajućim mjestima), dok se ostalih $l-r$ elemenata na početku i na kraju nalazi u nizu $2q+1$ puta (uz komplementarnost na odgovarajućim mjestima). Stoga, koristeći nezavisnost, funkciju S i lemu 2.5 dobivamo tvrdnju d).

Ideja dokaza tvrdnje d) može se primjeniti i na tvrdnju c) pazeći da sada u početnom i krajnjem dijelu nemamo elemente koji se pojavljuju $2q+3$ puta. □



Slika 2.2: Primjer povezanosti elemenata dva presijecajuća palindroma (puni krugovi predstavljaju elemente palindroma) duljine $m = 16$ za $2r > l$. Podebljane linije između elemenata predstavljaju sredine palindroma, a obične linije mjesta nužne simetrije.

Korolar 2.10 Pretpostavimo da su X_n nezavisne jednako distribuirane slučajne varijable te neka je $l = j - i$. Označimo s q i r odgovarajuće koeficijente kod cjelobrojnog dijeljenja s ostatkom $\frac{m}{2} = q \cdot l + r$. Tada vrijedi

a) za $\frac{m}{2} \leq l \leq m - 1$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x \in \mathcal{A}} p_x^2 p_{\widetilde{x}} \right)^{m-l} \left(\sum_{x \in \mathcal{A}} p_x p_{\widetilde{x}} \right)^{2l-m}$$

b) za $l < q$ i $2r \leq l$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x \in \mathcal{A}} p_x^{q+1} p_{\widetilde{x}}^{q+1} \right)^{2r} \left(\sum_{x \in \mathcal{A}} p_x^{q+1} p_{\widetilde{x}}^{q} \right)^{l-2r}$$

c) za $l < q$ i $2r > l$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x \in \mathcal{A}} p_x^{q+1} p_{\widetilde{x}}^{q+1} \right)^{2l-2r} \left(\sum_{x \in \mathcal{A}} p_x^{q+1} p_{\widetilde{x}}^{q+2} \right)^{2r-l}$$

Dokaz. Iz nezavisnosti i jednake distribuiranosti slučajnih varijabli X_n, napomene 2.6 te propozicije 2.9 slijedi tvrdnja korolara. □

Tvrdnju korolara 2.10 kao i dokaz može se naći i u radu Leung i dr. [9].

Napomena 2.11 Koristeći lemu 2.4 i propoziciju 2.9 (jednakosti 2.5 i 2.6 te korolar 2.10 u n.j.d. slučaju) kovarijanca $\mathrm{Cov}\,(Y_i, Y_j)$ može se izračunati budući da vrijedi

$$\mathrm{Cov}\,(Y_i, Y_j) = \mathrm{E}\,(Y_i \cdot Y_j) - \mathrm{E}\,(Y_i)\,\mathrm{E}\,(Y_j) = \mathrm{P}\,(Y_i \cdot Y_j = 1) - \mathrm{P}\,(Y_i = 1)\,\mathrm{P}\,(Y_j = 1) \quad (2.13)$$

■

Korolar 2.12 Pretpostavimo da su X_n nezavisne jednako distribuirane slučajne varijable s uniformnom distribucijom nad abecedom $\mathcal{A}$. Tada su pojavljivanja palindroma na mjestu

i i mjestu j nezavisna za $i \neq j$. Odnosno, vrijedi jednakost

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \mathrm{P}(Y_i = 1)\mathrm{P}(Y_j = 1) \ .$$

Dokaz. Ukoliko se palindromi ne sijeku nezavisnost je očita. Ako se pak sijeku, za $\beta = \frac{1}{|\mathcal{A}|}$ iz korolara 2.10 slijedi

a) za $\frac{m}{2} \leq l \leq m-1$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x\in\mathcal{A}} \beta^2 \cdot \beta\right)^{m-l} \left(\sum_{x\in\mathcal{A}} \beta \cdot \beta\right)^{2l-m} = \left(\beta^2\right)^{m-l} (\beta)^{2l-m} = \beta^m \ ,$$

b) za $l < q$ i $2r \leq l$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x\in\mathcal{A}} \beta^{q+1}\beta^{q+1}\right)^{2r} \left(\sum_{x\in\mathcal{A}} \beta^{q+1}\beta^{q}\right)^{l-2r} = \left(\beta^{2q+1}\right)^{2r} \left(\beta^{2q}\right)^{l-2r} = \beta^m \ ,$$

c) za $l < q$ i $2r > l$

$$\mathrm{P}(Y_i \cdot Y_j = 1) = \left(\sum_{x\in\mathcal{A}} \beta^{q+1}\beta^{q+1}\right)^{2l-2r} \left(\sum_{x\in\mathcal{A}} \beta^{q+1}\beta^{q+2}\right)^{2r-l}$$
$$= \left(\beta^{2q+1}\right)^{2l-2r} \left(\beta^{2q+2}\right)^{2r-l} = \beta^m \ .$$

Nezavisnost sada slijedi iz 2.9 budući da vrijedi

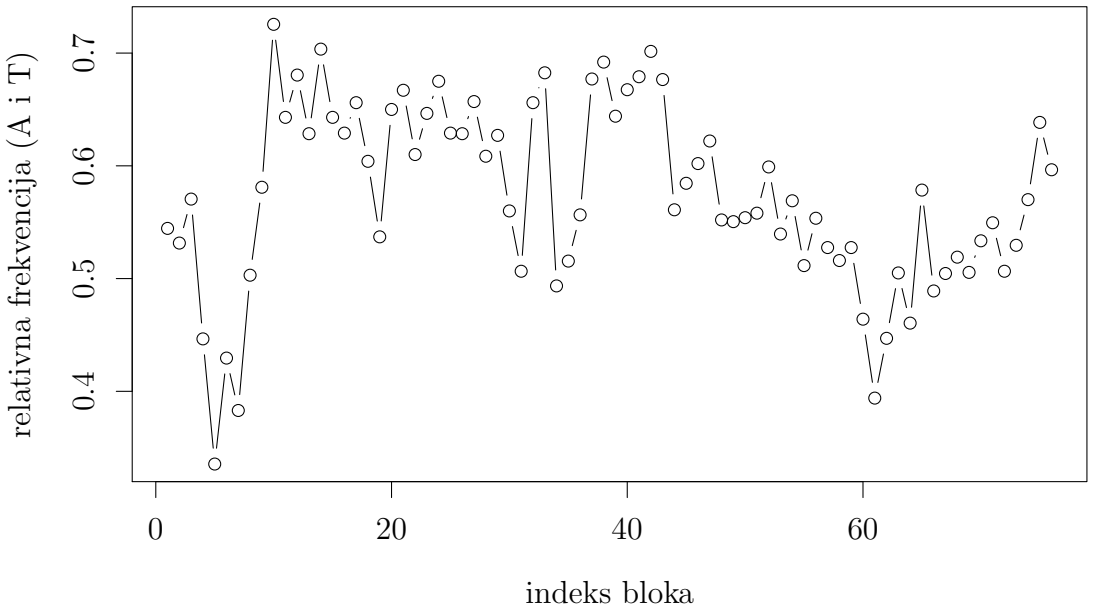
$$\mathrm{P}(Y_i = 1) = \prod_{j=1}^{\frac{m}{2}} \left[\sum_{x\in\mathcal{A}} \beta\beta\right] = \beta^{\frac{m}{2}} \ , \quad \mathrm{P}(Y_i = 1)\mathrm{P}(Y_j = 1) = \beta^{2\cdot\frac{m}{2}} = \beta^m \ .$$

□

2.4 Primjer primjene na stvarnim podacima

Iako dublja analiza stvarnih DNA nizova u odnosu na broj palindroma nije u području ovog doktorskog rada, svejedno će se u ovom potpoglavlju prezentirati kratka analiza određenog DNA niza kao primjer primjene na stvarnim podacima.

Poznata je činjenica da frekvencije nukleobaza variraju unutar DNA niza (Louie i dr. [14] te Mrazek i Karlin [17]). Ovo je svojstvo ilustrirano na slici 2.3 za niz „Homo sapiens chromosome 1 genomic contig NT_077912_1" [3] duljine 153 649. Niz je podijeljen na blokove jednake duljine od 2000 znakova (parova) te su zbrojene frekvencije komplementarnih baza A i T. Bitno je napomenuti da je duljina blokova odabrana proizvoljno i ne mora nužno reprezentirati smislenu razdiobu na regije. No, ukoliko se posjeduje ekspertno



Slika 2.3: Zbrojene frekvencije od A i T po blokovima duljine 2000 za niz „Homo sapiens chromosome 1 genomic contig NT_077912_1" [3]

znanje o nizu nad kojim se vrši istraživanje, blokove treba prilagoditi tom znanju. Ekspertno je znanje od velike važnosti za analizu budući da se kodirajuće i nekodirajuće regije uglavnom sastoje od različitih distribucija baza (Roy i dr. [24]).
Broj palindroma različitih duljina određen je po blokovima te je rezultat o asimptotskoj normalnoj distribuciji (pretpostavljajući da su blokovi tipa T1) primijenjen pri izračunu vjerojatnosti dobivanja broja palindroma minimalno velikog kao uočeni (p-vrijednosti ukoliko $\frac{N_n-\hat{\mu}_n}{\sqrt{n}}$ smatramo testnom statistikom). Rezultati za palindrome duljina 6 i 8 prezentirani su u tablici 2.2.

Tablica 2.2
Duljina niza $n = 153\,649$; za duljinu palindroma $m = 6$ bilo je 2464 palindroma dok je za duljinu palindroma $m = 8$ bilo 694 palindroma.

	$m = 6$			$m = 8$		
duljina blokova	$\frac{N_n-\hat{\mu}_n}{\sqrt{n}}$	$\hat{\sigma}^2$	p-vrijednost	$\frac{N_n-\hat{\mu}_n}{\sqrt{n}}$	$\hat{\sigma}^2$	p-vrijednost
400	-0.39555	0.01783	0.00153	0.0347	0.00458	0.3038
700	-0.41628	0.017767	0.00089	0.0319	0.00456	0.3181
1000	-0.42335	0.01772	0.00073	0.0322	0.00455	0.3162
1300	-0.44870	0.01775	0.00037	0.0234	0.00457	0.3646
2000	-0.45280	0.01770	0.00033	0.0236	0.00456	0.3633
n (n.j.d.)	-0.18075	0.01654	0.07995	0.12424	0.00422	0.02798

Zanimljivo je primijetiti nepodudaranje modela s blokovima s modelom u kojem pretpostavljamo nezavisnost i jednaku distribuiranost. Za duljinu palindroma $m = 6$ broj palindroma proglasili bi iznimnim, uz razinu značajnosti od 5%, za modele s blokovima dok istu vrijednost ne bi proglasili iznimnom po n.j.d. modelu. Obrnut slučaj možemo

primijetiti kod palindroma duljine $m = 8$ jer modeli s blokovima ne ukazuju na iznimnost za razliku od n.j.d. modela.

Ovaj primjer tako pokazuje da bi korištenje n.j.d. modela moglo dovesti do krivog tumačenja. Razlog leži u regijama s visokim očekivanim brojem palindroma (zbog visokog postotka komplementarnih baza) pa korištenjem n.j.d. modela gubimo tu informaciju budući da usrednjavamo frekvencije. Slobodno govoreći, model s blokovima je sposoban prepoznati razlike među regijama dok n.j.d. model to isto nije sposoban prepoznati. Stoga, n.j.d. model treba koristiti s oprezom.

Također, zanimljivo je primijetiti da se promjenom veličine blokova p-vrijednosti ne mijenjaju znatno.

Dodatak A

C++ kôd za optimalno poravnanje Damerau-Levenshteinovim algoritmom

Kôd prezentiran ovdje jest Proof-of-concept te se zasniva na rekurzivnoj definiciji ne mareći za efikasnost.

```
0   #include <vector>
1   typedef std::vector<int> rijec;
2
3   rijec ptraj(int c, rijec t) {
4       rijec nova;
5       nova.push_back(c);
6       nova.insert(nova.end(), t.begin(), t.end());
7       return nova;
8   }
9
10  rijec ptraj(rijec c, rijec t) {
11      c.insert(c.end(), t.begin(), t.end());
12      return c;
13  }
14
15  struct levRez {
16      int rez;
17      std::vector<rijec> traj;
18
19      levRez() {}
20  };
21
22  // MATCH=1, REPLACE=2, INSERT=3,
23  // DELETE=4, TRANSPOSITION=5
```

```
24  levRez damlev(rijec str1, rijec str2, int i, int j, rijec traj) {
25      levRez lr;
26      if (std::min(i, j) == 0) {
27          if (i > 0)
28              traj = ptraj(rijec(i, 4), traj);
29          else if (j > 0)
30              traj = ptraj(rijec(j, 3), traj);
31          lr.rez = std::max(i, j);
32          lr.traj.push_back(traj);
33          return lr;
34      }
35
36      int x = 0;
37      rijec tmp;
38      if (str1[i - 1] !=
39          str2[j - 1]) { // i-1 sluzi prilagodbi indeksiranju od nule
40          x = 1;
41          tmp = ptraj(2, traj);
42      } else
43          tmp = ptraj(1, traj);
44
45      std::vector<levRez> L;
46      L.push_back(damlev(str1, str2, i - 1, j, ptraj(4, traj)));
47      L.push_back(damlev(str1, str2, i, j - 1, ptraj(3, traj)));
48      L.push_back(damlev(str1, str2, i - 1, j - 1, tmp));
49
50      int transp = 0;
51      if (i > 2 && j > 2 && str1[i - 2] == str2[j - 1] &&
52          str1[i - 1] == str2[j - 2]) {
53          L.push_back(damlev(str1, str2, i - 2, j - 2, ptraj(5, traj)));
54          transp = 1;
55      }
56
57      L[0].rez += 1;
58      L[1].rez += 1;
59      L[2].rez += x;
60      int m = std::min(L[0].rez, L[1].rez);
61      m = std::min(m, L[2].rez);
62      lr.rez = m;
```

```
63
64      if (transp) {
65          L[3].rez += 1;
66          lr.rez = std::min(lr.rez, L[3].rez);
67      }
68
69      // upisi minove u lr
70      for (auto &x : L) {
71          if (x.rez == lr.rez) {
72              for (auto &y : x.traj) {
73                  lr.traj.push_back(y);
74              }
75          }
76      }
77
78      return lr;
79  }
```

Bibliografija

[1] Hock Peng Chan, Nancy Ruonan Zhang i Louis H.Y. Chen. "Importance Sampling of Word Patterns in DNA and Protein Sequences". *Journal of Computational Biology* 17.12 (prosinac 2010.), str. 1697–1709. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2008.0233.

[2] David S. H. Chew, Kwok Pui Choi i Ming-Ying Leung. "Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses". *Nucleic Acids Research* 33.15 (2005.), e134. ISSN: 0305-1048. DOI: 10.1093/nar/gni135.

[3] NCBI database. *Homo sapiens (human) (Annotation Release 104)*.

[4] Thomas Shelburne Ferguson. *A Course in Large Sample Theory*. en. CRC Press, 1996. ISBN: 9780412043710.

[5] Stéphane Fischler. "Palindromic prefixes and episturmian words". *Journal of Combinatorial Theory, Series A* 113.7 (listopad 2006.), str. 1281–1304. ISSN: 0097-3165. DOI: 10.1016/j.jcta.2005.12.001.

[6] Ranko Gaćeša. "Zastupljenost palindroma u kodirajućoj DNA kvasca Saccharomyces cerevisiae". Croatian. Mag. rad. Prehrambeno-biotehnološki fakultet, Sveučilište u Zagrebu, 2011.

[7] Joao P. Gomes, William J. Bruno, Alexandra Nunes, Nicole Santos, Carlos Florindo, Maria J. Borrego i Deborah Dean. "Evolution of Chlamydia trachomatis diversity occurs by widespread interstrain recombination involving hotspots". *Genome Research* 17.1 (siječanj 2007.), str. 50–60. ISSN: 1088-9051. DOI: 10.1101/gr.5674706.

[8] Allan Gut. *Probability: A Graduate Course*. Springer texts in statistics. Springer, 2012. ISBN: 9781461447078.

[9] Ming-Ying Leung, Kwok Pui Choi, Aihua Xia i Louis H.Y. Chen. "Nonrandom Clusters of Palindromes in Herpesvirus Genomes". *Journal of Computational Biology* 12.3 (travanj 2005.), str. 331–354. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2005.12.331.

[10] Berislav Lisnić, Ivan-Kresimir Svetec, Hrvoje Sarić, Ivan Nikolić i Zoran Zgaga. "Palindrome content of the yeast Saccharomyces cerevisiae genome". eng. *Current genetics* 47.5 (svibanj 2005.), str. 289–297. ISSN: 0172-8083. DOI: 10.1007/s00294-005-0573-5.

[11] Kirill S Lobachev, Alison Rattray i Vidhya Narayanan. "Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells". eng. *Frontiers in bioscience: a journal and virtual library* 12 (2007.), str. 4208–4220. ISSN: 1093-4715.

[12] M Lothaire. *Applied combinatorics on words*. English. Encyclopedia of Mathematics and its Applications. Cambridge, UK; New York: Cambridge University Press, 2005. ISBN: 9781461938316 1461938317 9781107341005 1107341000.

[13] M. Lothaire. "Statistics on Words with Applications to Biological Sequences". *Applied Combinatorics on Words*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2005. ISBN: 9781107341005.

[14] Elizabeth Louie, Jurg Ott i Jacek Majewski. "Nucleotide Frequency Variation Across Human Genes". *Genome Research* 13.12 (prosinac 2003.), str. 2594–2601. ISSN: 1088-9051. DOI: 10.1101/gr.1317703.

[15] Le Lu, Hui Jia, Peter Dröge i Jinming Li. "The human genome-wide distribution of DNA palindromes". eng. *Functional & integrative genomics* 7.3 (srpanj 2007.), str. 221–227. ISSN: 1438-793X. DOI: 10.1007/s10142-007-0047-6.

[16] C. Matek, T. E. Ouldridge, J. P. Doye i A. A. Louis. "Studying molecular physiology of DNA cruciforms with a coarse-grained computational model". *Proceedings of The Physiological Society*. 37th Congress of IUPS. 2013.

[17] Jan Mrazek i Samuel Karlin. "Strand compositional asymmetry in bacterial and large viral genomes". *Proceedings of the National Academy of Sciences of the United States of America* 95.7 (ožujak 1998.), str. 3720–3725. ISSN: 0027-8424.

[18] D K Nag i A Kurst. "A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast Saccharomyces cerevisiae". eng. *Genetics* 146.3 (srpanj 1997.), str. 835–847. ISSN: 0016-6731.

[19] Steven Orey. "A central limit theorem for m-dependent random variables". *Duke Mathematical Journal* 25.4 (prosinac 1958.), str. 543–546. ISSN: 0012-7094. DOI: 10.1215/S0012-7094-58-02548-1.

[20] Željko Pauše. *Uvod u matematičku statistiku*. Školska knjiga, 1993.

[21] Hrvatski jezični portal. URL: http://hjp.novi-liber.hr/ (pogledano 17.12.2013.).

[22] Stéphane Robin, François Rodolphe i Sophie Schbath. *DNA, Words and Models: Statistics of Exceptional Words*. Translated from French. Cambridge, UK; New York: Cambridge University Press, 2005. ISBN: 052184729X 9780521847292.

[23] Stéphane Robin i Sophie Schbath. "Numerical Comparison of Several Approximations of the Word Count Distribution in Random Sequences". *Journal of Computational Biology* 8.4 (rujan 2001.), str. 349–359. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/106652701752236179.

[24] M. Roy, S. Biswas i S. Barman. "Identification and analysis of coding and non-coding regions of a DNA sequence by positional frequency distribution of nucleotides (PFDN) algorithm". *4th International Conference on Computers and Devices for Communication, 2009. CODEC 2009*. 2009., str. 1–4.

[25] N. Sarapa. *Teorija vjerojatnosti*. Školska knjiga, 2002. ISBN: 9789530308169.

[26] Mark J. Schervish. *Theory of Statistics*. en. Springer, kolovoz 1995. ISBN: 9780387945460.

[27] A. Shiryaev i S.S. Wilson. *Probability*. Graduate Texts in Mathematics. Springer, 1995. ISBN: 9780387945491.

[28] R R Sinden, S S Broyles i D E Pettijohn. "Perfect palindromic lac operator DNA sequence exists as a stable cruciform structure in supercoiled DNA in vitro but not in vivo". eng. *Proceedings of the National Academy of Sciences of the United States of America* 80.7 (travanj 1983.), str. 1797–1801. ISSN: 0027-8424.

[29] Drago Špoljarić i Ivo Ugrina. "On Statistical Properties of Palindromes in DNA". *Communications in Statistics - Theory and Methods* 42.7 (2013.), str. 1373–1385. ISSN: 0361-0926. DOI: 10.1080/03610926.2012.739253.

[30] S K Thukral, A Eisen i E T Young. "Two monomers of yeast transcription factor ADR1 bind a palindromic sequence symmetrically to activate ADH2 expression". eng. *Molecular and cellular biology* 11.3 (ožujak 1991.), str. 1566–1577. ISSN: 0270-7306.

[31] A S Waldman, H Tran, E C Goldsmith i M A Resnick. "Long inverted repeats are an at-risk motif for recombination in mammalian cells." *Genetics* 153.4 (prosinac 1999.), str. 1873–1883. ISSN: 0016-6731.

[32] Zhiyuan Zhai, Gesine Reinert, Kai Song, Michael S. Waterman, Yihui Luan i Fengzhu Sun. "Normal and Compound Poisson Approximations for Pattern Occurrences in NGS Reads". *Journal of Computational Biology* 19.6 (lipanj 2012.), str. 839–854. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0029.

Kazalo pojmova

Popis odabranih stručnih pojmova s odgovarajućim engleskim prijevodom.

Indeks pojmova

Popis slika

Životopis

Ivo Ugrina rođen je 21. svibnja 1983. godine u Splitu u Hrvatskoj. U Splitu završava osnovnu i srednju školu te potom 2001. godine upisuje Prirodoslovno-matematički fakultet u Zagrebu. Za vrijeme studija obavlja dužnost demonstratora iz nekoliko kolegija na inženjerskom smjeru matematike. U rujnu 2008. godine uspješno brani diplomski rad stekavši time titulu diplomiranog inženjera matematike. Akademske godine 2008./2009. upisuje doktorski studij matematike pri Prirodoslovno-matematičkom fakultetu u Zagrebu gdje se i zapošljava kao asistent na projektu „Mireo World" u veljači 2009. godine. Kao asistent u nastavi sudjelovao je u izvođenju vježbi iz nekoliko računarskih i matematičkih kolegija. Tijekom izobrazbe za doktora znanosti sudjelovao je u radu seminara za Teoriju vjerojatnosti i matematičku statistiku.

Objavio je dva rada u SCIE referentnim časopisima te sudjelovao na desetak konferencija, kako stručnih tako i znanstvenih.

Bibliografija

1. Drago Špoljarić i Ivo Ugrina. "On Statistical Properties of Palindromes in DNA".: *Communications in Statistics - Theory and Methods* 42.7 (2013.), str. 1373–1385. ISSN: 0361-0926. DOI: 10.1080/03610926.2012.739253
2. Sanja Perković, Sandra Bašić-Kinda, Igor Aurer, Ivo Ugrina, Antica Duletić-Načinović, Dominik Lozić i Drago Batinić. "Multiparameter flow cytometry is necessary for detection, characterization and diagnostics of composite mature B-cell lymphoproliferative neoplasms". en. *International Journal of Hematology* 98.5 (studeni 2013.), str. 589–596. ISSN: 0925-5710, 1865-3774. DOI: 10.1007/s12185-013-1432-7