Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kristjan Vimm

Suboptimaalsed meetodid jadade sarnasuse võrdlemiseks kahetähelise tähestiku puhul

Matemaatika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Jüri Lember

Tartu 2022

SUBOPTIMAALSED MEETODID JADADE SARNASUSE VÕRDLEMISEKS KAHETÄHELISE TÄHESTIKU PUHUL

Bakalaureusetöö

Kristjan Vimm

Lühikokkuvõte

Töös kirjeldatakse ja võrreldakse kolme suboptimaalset meetodit jadade joondamiseks ja nende sarnasuse mõõtmiseks kahetähelise tähestiku puhul. Defineeritakse joondus, skoor, pikim ühisjada, lõpmatute jadade sõltumatus ja *i.i.d.* jada. Võetakse kasutusele jadade blokkide kaupa esitus, blokipikkuste jada, blokipaaride pikkuste jada, blokipaaride skooride jada ning tõestatakse nende olulisemad omadused. Seejärel tõestatakse Hammingu joonduse, bloki kaupa Hammingu joonduse ja järjestikuse joonduse keskmiste skooride piirväärtuste koondumised vastavateks konstantideks. Lõpetuseks koostatakse simulatsioon, võrreldakse teoreetilisi tulemusi simulatsiooni tulemustega ning võrreldakse välja arvutatud pikima ühisjada pikkust iga meetodi skooriga.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Jadade sarnasus, skoor, pikim ühisjada, Hammingu joondus.

SUBOPTIMAL METHODS FOR COMPARING THE SIMILARITY OF SEQUENCES FOR A TWO-LETTER ALPHABET

Bachelor thesis

Kristjan Vimm

Abstract

This thesis focuses on three suboptimal methods of aligning sequences and measuring their similarity for a two-letter alphabet. The definitions of alignment, score, longest common subsequence, independence of infinite sequences and *i.i.d.* sequence are given. Tools like the block representation of a sequence, block length sequence, block-pair length sequence and block-pair score sequence are implemented and their most important properties are proven. Then the convergence of the limits of the average scores of Hamming's alignment, block-by-block Hamming's alignment and consecutive alignment are proven. Finally, a simulation is conducted and the results for every alignment are compared to their respective theoretical results and also to the longest common subsequence calculated during the simulation.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Similarity of sequences, score, longest common subsequence, Hamming's alignment.

Sisukord

Sissejuhatus

Mitmes valdkonnas on oluliseks uurimisobjektiks jadad ning nendevaheline sarnasus. Seetõttu on loodud üldine teooria suvaliste jadade sarnasuse võrdlemiseks ning meetodid nagu näiteks Needleman-Wunschi dünaamilise planeerimise algoritm, mis suudavad seda sarnasust täpselt hinnata. Need on laialdaselt kasutatavad meetodid, aga pikkade jadade korral on nende töömaht äärmiselt suur, mistõttu on uuritud ka suboptimaalseid meetodeid, mis ei anna küll täpset infot jadade sarnasuse kohta, kuid toimivad palju kiiremini. (Barder *et al.*, 2012)

Selles töös võrdleme Bernoulli jaotusega juhuslikest suurustest koosnevaid jadasid $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$ ning vaatleme kolme suboptimaalset meetodit. Täpsemalt uurime, milline nendest kolmest meetodist annab kõige parema tulemuse Bernoulli jaotuse parameetrite p^x ja p^y muutudes ning kas leidub meetod, mis ei ole kunagi parem kui teised.

Esimeses peatükis tutvustame jadade võrdlemise üldisi mõisteid, võtame kasutusele abivahendid vaadeldavate meetodite uurimiseks ning tõestame iga meetodi jaoks koondumise tema keskmise skoori piirväärtuseks. Mainitud abivahendid peavad rahuldama kindlaid tingimusi, mille tõestamiseks kulub suur osa esimesest peatükist. Teise peatüki alguses üritame ennustada iga meetodi käitumist kindlatel erijuhtudel ning võrdleme oma ennustusi teoreetiliste tulemustega. Lõpetuseks viime läbi simulatsiooni, mis võimaldab meil juhuslikult genereeritud jadade jaoks välja arvutada täpse sarnasuse skoori ning võrrelda seda suboptimaalsete meetodite skooridega.

1 Teoreetiline taust

Selles peatükis tutvustame üldisi mõisteid, mida läheb vaja jadade võrdlemiseks ning võtame kasutusele ka spetsiifilised abivahendid nende jadade võrdlemise meetodite jaoks, mida hakkame uurima.

1.1 Determineeritud jadad

Selle alapeatüki definitsioonid ning arutelu on kirjutatud bakalaureusetöö Toots (2008) põhjal.

Definitsioon 1. *Lõplikku hulka Σ nimetatakse tähestikuks ning selle elemente nimetatakse tähtedeks.*

Näide. Selles töös kasutame *Bernoulli* jaotusega jadasid, mille elemendid on tähestikust

$$\Sigma = \{0, 1\}.$$

Definitsioon 2. *Olgu $n \in \mathbb{N}$ fikseeritud naturaalarv. Hulga*

$$\Sigma^n = \{(x_1, x_2, \ldots, x_n) : x_i \in \Sigma \text{ ja } 1 \leq i \leq n\}$$

elementi nimetatakse n-elemendiliseks jadaks ning tähistatakse $x = x_1x_2 \ldots x_n$ või $x = x_1, x_2, \ldots, x_n$.

Näide. $x = 01100100$ on jada tähestikul $\Sigma = \{0, 1\}$ ning $x \in \Sigma^8$.

Definitsioon 3. *Olgu $x \in \Sigma^K$ mingi lõplik või lõpmatu jada tähestikul Σ, kus $K \in \mathbb{N} \cup \{\infty\}$. Jada $y \in \Sigma^R$, kus $R \leq K$, nimetame jada x osajadaks, kui leidub hulk*

$$\{n_1, n_2, \ldots, n_R\} \subseteq \{1, 2, \ldots, K\} \text{ nii, et}$$

$$1 \leq n_1 < n_2 < \ldots < n_R \leq K \text{ ja } y_i = x_{n_i}, \forall i \in 1, 2, \ldots, R \text{ korral}$$

Näide. Jada $x =$ ABCDEFGH osajadad on muuhulgas AB, BC, ABGH, ACEG. Võib mõelda, et osajada saadakse algse jada mingi (kasvõi tühja) hulga elementide ärajätmisel, aga elementide järjekord peab jääma samaks.

Definitsioon 4. *Olgu $x \in \Sigma^K$ ja $y \in \Sigma^L$ mingid lõplikud või lõpmatud jadad tähestikul Σ, kus $K \in \mathbb{N} \cup \{\infty\}$ ja $L \in \mathbb{N} \cup \{\infty\}$. Jada $z \in \Sigma^R$, kus $R \leq min(K, L)$, nimetatakse jadade x ja y ühisjadaks, kui leiduvad hulgad*

$$\{n_1, n_2 \ldots, n_R\} \subseteq \{1, 2, \ldots, K\} \text{ ja}$$

$$\{m_1, m_2 \ldots, m_R\} \subseteq \{1, 2, \ldots, L\} \text{ nii, et}$$

$$1 \leq n_1 < n_2 < \ldots < n_R \leq K,\ 1 \leq m_1 < m_2 < \ldots < m_R \leq L \text{ ja}$$

$$z_i = x_{n_i} = y_{m_i},\ \forall i \in 1, 2, \ldots, R \text{ korral.}$$

Näide. Tähestiku $\Sigma = \{\text{A,T,C,G}\}$ korral on jadadel $x = \text{AGGCATA}$ ja $y = \text{ACCAGGTA}$, kus $x \in \Sigma^7, y \in \Sigma^8$, üheks ühisjadaks jada $z = \text{ACT}$, sest leiduvad indeksid

$$n_1 = 1, n_2 = 4, n_3 = 6 \text{ ja } m_1 = 1, m_2 = 3, m_3 = 7 \text{ nii, et}$$

$$1 \leq n_1 < n_2 < n_3 \leq 7,\ 1 \leq m_1 < m_2 < m_3 \leq 8 \text{ ning}$$

$$z_1 = x_{n_1} = y_{m_1} = \text{A},\ z_2 = x_{n_2} = y_{m_2} = \text{C ja } z_3 = x_{n_3} = y_{m_3} = \text{T}.$$

Definitsioon 5. *Jadade x ja y pikimaks ühisjadaks nimetatakse sellist x ja y ühisjada, mille pikkus on maksimaalne.*

Näide. Kasutades eelmise näite jadasid näeme, et jada $z = \text{AGGTA}$ on jadade x ja y pikim ühisjada, sest mõlema jada puhul saame valida indeksid nii, et moodustub osajada z ehk z on nende ühisjada ning ei leidu x ja y ühisjada, mis kuuluks hulka Σ^6 või Σ^7. Paneme tähele, et pikim ühisjada ei pruugi olla üheselt määratud. Jadadel x ja y leidub ka pikim ühisjada $w = \text{ACATA}$. Neid ühisjadasid saab visualiseerida järgnevalt:

$$\begin{array}{ll} \text{A}---\text{GGCATA} & \text{AGG}-\text{CA}--\text{TA} \\ \text{ACCAGG}--\text{TA} & \text{A}--\text{CCAGGTA} \end{array}$$

Siin võtsime juba kasutusele sümboli $-$, mida nimetatakse *indeliks*. See tuleneb ingliskeelsetest sõnadest *insertion* ja *deletion* ehk sisestamine ja kustutamine, mis väljendavad seda, kuidas näiteks DNA ahelas võib mutatsioonide tõttu mõni täht vahelt ära jääda või juurde tekkida.

Võtame kasutusele mõned definitsioonid sellise visualiseerimise vormistamiseks.

Definitsioon 6. *Tähestiku Σ laienduseks ehk laiendatud tähestikuks nimetatakse tähestikku $\Sigma_+ = \Sigma \cup \{-\}$.*

Definitsioon 7. *Olgu meil jada $x \in \Sigma^K$. Siis jada x laiendatud jadaks nimetatakse jada $y \in \Sigma_+^L$, kui $K \leq L$ ning leidub hulk*

$$\{n_1, n_2, \ldots, n_K\} \subset \{1, 2, \ldots, L\} \textit{ nii, et}$$

$$1 \leq n_1 < n_2 < \ldots < n_K \leq L \textit{ ja } x_i = y_{n_i}, \forall i \in 1, 2, \ldots, K \textit{ korral ja}$$

$$y_i = -, \forall j \in \{1, 2, \ldots, L\} \setminus \{n_1, n_2, \ldots, n_K\}$$

Näide. Toome mõned jada $x = \text{ABCDEFG}$ laiendatud jadade näited:

- A$-$BCDEFG
- $---$ABCD$-$E$-$F$-$G$--$

- $---$ABCDEFG$---$.

Definitsioon 8. *Olgu x ja y jadad. Paari (x^*, y^*) nimetatakse jadade x ja y joonduseks, kui*

1. *x^* on jada x ja y^* on jada y laiendatud jada.*
2. *x^* ja y^* on mõlemad pikkusega $n \in \mathbb{N}$.*
3. *$\forall i \in 1, 2, \ldots, n : \neg[(x_i^* = -) \ \wedge \ (y_i^* = -)]$, s.t kaks indelit ei asu kunagi vastakuti.*

Näide. Jadade $x =$ ABCDEF ja $y =$ GHIJKL joondus on näiteks

$$\begin{matrix} \text{AB}-\text{CDE}-\text{F} \\ \text{G}-\text{H}-\text{I JKL} \end{matrix} \text{, aga mitte } \begin{matrix} -\text{ABCDEF} \\ -\text{GH I JK}-\text{L} \end{matrix}$$

Võtame kasutusele kaks laensõna inglise keelest, et kirjeldada joonduses esinevaid olukordi. Kui mingi $i \in \{1, 2, \ldots, n\}$ korral on kohakuti asuvad tähed laiendatud jadades x^* ja y^* võrdsed ehk $x_i^* = y_i^*$, siis ütleme selle kohta *match*. Kui aga kohakuti asuvad tähed on erinevad ning kumbki neist pole indel ehk $- \neq x_i^* \neq y_i^* \neq -$, siis ütleme selle kohta *mismatch*.

Definitsioon 9. *Olgu Σ tähestik ja Σ_+ tema laiendatud tähestik. Funktsiooni $s : \Sigma_+ \times \Sigma_+ \to \mathbb{R}$ nimetame sarnasusfunktsiooniks.*

Olgu $\mu, \delta \in \mathbb{R}, \mu \geq 0, \delta \geq 0$. Üks tüüpiline viis sarnasusfunktsiooni defineerida on

$$s(x_i^*, y_i^*) = \begin{cases} 1, & \text{kui } x_i^* = y_i^* \ (match) \\ -\mu, & \text{kui } - \neq x_i^* \neq y_i^* \neq - \ (mismatch) \\ -\delta, & \text{kui } x_i^* = - \text{ või } y_i^* = - \end{cases}.$$

Definitsioon 10. *Olgu meil jadad x ja y. Joonduse (x^*, y^*) skoori defineerime järgnevalt*

$$s(x^*, y^*) = \sum_{i=1}^{n} s(x_i^*, y_i^*), \ \textit{kus } n \textit{ on joonduse } (x^*, y^*) \textit{ pikkus.}$$

Skoor aitab kvantifitseerida seda, kui sarnased või erinevad on võrreldavad jadad lähtudes valitud sarnasusfunktsioonist. Sarnasust väljendab kõige paremini optimaalne skoor.

Definitsioon 11. *Kahe jada x ja y optimaalseks skooriks $S = S(x, y)$ nimetatakse maksimaalset skoori üle kõigi joonduste:*

$$S(x, y) = \max\{s(x^*, y^*) : (x^*, y^*) \textit{ on jadade } x \textit{ ja } y \textit{ joondus}\}.$$

Joondust, mis maksimeerib skoori, nimetame optimaalseks joonduseks.

Näide. Kasutame ühe varasema näite jadasid $x = \text{AGGCATA}$ ja $y = \text{ACCAGGTA}$ ning nende ühte võimalikku joondust

$$\begin{array}{c} \text{A}--\text{GG}-\text{CATA} \\ \text{ACCAGG}--\text{TA} \end{array}$$

Tüüpilise sarnasusfunktsiooni põhjal on selle joonduse skooriks

$$\begin{aligned} s(x^*, y^*) = {} & s(A,A) + s(-,C) + s(-,C) + s(G,A) + s(G,G) + \\ & s(-,G) + s(C,-) + s(A,-) + s(T,T) + s(A,A) = 4 - \mu - 5\delta. \end{aligned}$$

Loomulikult sõltub sellist tüüpi sarnasusfunktsiooniga leitud skoor konstantide $\mu \geq 0$ ja $\delta \geq 0$ valikust. Üks eriline juht on sarnasusfunktsioon, kus $\mu = 1$ ning $\delta = 0$ ehk *mismatch* on sama kaaluga kui *match*, aga negatiivse väärtusega ning *indel* on n-ö tasuta. Sellisel juhul ei sisalda ükski optimaalne joondus mitte ühtegi *mismatch*'i ning optimaalne skoor on ühtlasi ka pikima ühisjada pikkus. Selle nägemiseks eeldame, et leidub optimaalne joondus, mis sisaldab *mismatch*'i ehk selles leiduvad kaks *indel*'ist erinevat tähte, mis on kohakuti. Olgu need B ja C. Lisame kummalegi jadale ühe *indel*'i näiteks nii:

$$\begin{array}{c} \text{AABAA} \\ \text{AACAA} \end{array} \rightarrow \begin{array}{c} \text{AA}-\text{BAA} \\ \text{AAC}-\text{AA} \end{array}$$

Sellise muudatuse korral suureneb joonduse skoor ühe võrra, sest me asendasime ühe *mismatch*'i kahe *indel*'iga. Jõudsime vastuoluni eeldusega, et joondus oli optimaalne, seega kehtib, et sellise sarnasusfunktsiooni puhul ei leidu üheski optimaalses joonduses *mismatch*'e. Siit järeldub, et kui a on *match*'ide arv, b on *mismatch*'ide arv ning c on *indel*'ite arv, siis sellise sarnasusfunktsiooni puhul on optimaalne skoor alati

$$S(x^*, y^*) = a \cdot 1 - b \cdot \mu - c \cdot \delta = a - 0 \cdot 1 - c \cdot 0 = a.$$

Kuna a peab olema optimaalne, siis see on suurim võimalik *match*'ide arv ehk ka pikima ühisjada pikkus. Märgime, et optimaalne skoor on alati üheselt määratud nagu ka pikima ühisjada pikkus, aga optimaalne joondus, mis annab vastava optimaalse skoori, ei pruugi olla üheselt määratud nagu ka pikim ühisjada ise.

Edaspidi kasutamegi sarnasusfunktsiooni

$$s(x_i^*, y_i^*) = \begin{cases} 1, & \text{kui } x_i^* = y_i^* \ (match) \\ -1, & \text{kui } - \neq x_i^* \neq y_i^* \neq - \ (mismatch) \\ 0, & \text{kui } x_i^* = - \text{ või } y_i^* = - \end{cases},$$

seega optimaalse joonduse leidmiseks tuleb vaid leida joondus, mis moodustab ühe pikima ühisjada. See on kasulik, sest hiljem tahame teada, kas vaadeldavad suboptimaalsed joondused annavad häid skoore ja selleks võrdleme neid optimaalse skooriga, milleks piisab selle sarnasusfunktsiooni korral leida pikima ühisjada pikkus.

1.2 Juhuslike suuruste jadad

Enne kui saame suboptimaalsete meetodite skoore omavahel võrrelda, tuleb täpsustada, millised on need kaks juhuslike suuruste jada X ja Y, mida me omavahel võrdleme ning tõestada tulemusi, mis aitavad hinnata nende meetodite skoore.

Edaspidi vaatleme lõpmatuid juhuslike suuruste jadasid tähestikul $\Sigma = \{0, 1\}$. See tähendab, et vaadeldavad jadad on $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$, mille liikmed on Bernoulli jaotusega, kusjuures $X(n)$ tähistab jada X esimest n liiget.

Definitsioon 12. *Lõpmatu jada $X = X_1, X_2\ldots$ on i.i.d. jada (independent and identically distributed), kui kõik jada X liikmed on*

- *sama jaotusega, s.t $\forall i, j, k \in \mathbb{N}, P(X_i = k) = P(X_j = k)$ ja*
- *omavahel sõltumatud s.t $\forall n, k_1, \ldots, k_n$,*

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) = P(X_1 = k_1)\ldots P(X_n = k_n).$$

Lõplik jada $X(n) = X_1, \ldots, X_n$ on i.i.d. jada, kui kõik jada $X(n)$ liikmed on sama jaotusega ja omavahel sõltumatud.

Jada X sama jaotuse ja sõltumatuse tingimustega on samaväärne tingimus $\forall n, k_1$, $\ldots, k_n$

$$P(X_1 = k_1, \ldots, X_n = k_n) = P(X_1 = k_1)P(X_1 = k_2)\ldots P(X_1 = k_n),$$

sest sama jaotuse omaduse abil saame sõltumatuse tingimuses iga $X_i, i = 1, 2, \ldots, n$ asendada juhusliku suurusega X_1.

Definitsioon 13. *Ütleme, et lõpmatud jadad $X = X_1, X_2, \ldots$ ja $Y = Y_1, Y_2, \ldots$ on omavahel sõltumatud, kui kõigi naturaalarvude $n, k_1, \ldots, k_n, l_1, \ldots, l_n$ korral kehtib võrdus*

$$P(X_1 = k_1, \ldots, X_n = k_n, Y_1 = l_1, \ldots, Y_n = l_n) =$$
$$= P(X_1 = k_1)\ldots P(X_n = k_n)P(Y_1 = l_1)\ldots P(Y_n = l_n).$$

Olgu meie lõpmatud, Bernoulli jaotusega juhuslike suuruste jadad X ja Y ka omavahel sõltumatud *i.i.d.* jadad, kus $P(X_1 = 1) = p^x,\ P(X_1 = 0) = 1 - p^x,\ P(Y_1 = 1) = p^y,\ P(Y_1 = 0) = 1 - p^y$. Piisab välja tuua ainult esimeste liikmete X_1 ja Y_1 jaotused, sest X_1 on sama jaotusega, mis X_i, $\forall i \in 1, \ldots, n$ ning samuti ka jada Y puhul.

Kaks joonduse meetodit, mida me edaspidi uurime, ei joonda jadasid ainult tähtede kaupa, vaid koondavad järjestikused võrdsed tähed blokkidesse ning joondavad saadud blokke omavahel või kasutavad jada blokkide esitust skoori leidmiseks.

Näide. Vaatleme jada $x = 0011101111000$. Loomulik viis jaotada see jada blokkideks oleks nii, et esimene on nullide blokk pikkusega 2, teine on ühtede blokk pikkusega 3, kolmas on nullide blokk pikkusega 1, neljas on ühtede blokk pikkusega 4 ning viimane on nullide blokk pikkusega 3. Kui tahaksime seda jada võrrelda jadaga $y = 10000110011$, mille blokkide esitus on 10101 ning blokkide pikkused on vastavalt 1, 4, 2, 2 ja 2, siis võiksime valida blokkide kaupa joonduse

$$\begin{matrix} -01010 \\ 10101- \end{matrix}',$$

millele vastav tähtede joondus on olenevalt *indel*'ite paigutusest näiteks

$$\begin{matrix} -00--1110-1111000 \\ 1000011-0011----- \end{matrix} \text{ või } \begin{matrix} ---00111-01111000 \\ 10000-1100--11--- \end{matrix}.$$

Järgmine definitsioon on võetud magistritööst Toots (2012) ning kohandatud praeguseks juhuks.

Definitsioon 14. *Olgu meil jada $X(n) = X_1, X_2, \ldots, X_n$ jaoks antud sellised indeksid $0 =: i_0 < i_1 < i_2 < \ldots < i_m := n$, et*

- *iga $l \in \{1, \ldots, m\}$ korral leidub selline $v \in \Sigma$, et kui $i \in \{i_{l-1} + 1, \ldots, i_l\}$, siis $X_i = v$, s.t indeksid $i_0 < i_1 < \ldots < i_m$ jaotavad jada X blokkideks,*
- *iga $l \in \{1, \ldots, m-1\}$ korral kehtib $X_{i_l} \neq X_{i_{l+1}}$, s.t kaks järjestikust blokki ei koosne samadest tähtedest.*

Jada

$$L^x = L^x_1, L^x_2, \ldots, L^x_m = i_1 - i_0, i_2 - i_1, \ldots, i_m - i_{m-1}$$

nimetame jada X bloki pikkuste jadaks.

Näide. Täpsustame eelmise näite jada $x = 0011101111000$ indeksid uue definitsiooni jaoks:

$$i_0 = 0,\ i_1 = 2,\ i_2 = 5,\ i_3 = 6,\ i_4 = 10,\ i_5 = 13.$$

Blokkide arv on $m = 5$ ning bloki pikkuste jada on $L = 23143$.

Definitsioon 15. *Olgu meil antud jadad $X(n)$, $Y(n)$ ja nende bloki pikkuste jadad $L^x(t) = L^x_1 L^x_2 \ldots L^x_t$ ja $L^y(s) = L^y_1 L^y_2 \ldots L^y_s$. Olgu*

$$i := \begin{cases} t/2, & \text{kui } t \text{ on paarisarv,} \\ (t-1)/2, & \text{kui } t \text{ on paaritu} \end{cases} \text{ ja } j := \begin{cases} s/2, & \text{kui } s \text{ on paarisarv,} \\ (s-1)/2, & \text{kui } s \text{ on paaritu arv.} \end{cases}$$

Jada

$$Z^x(i) = Z_1^x, Z_2^x, \ldots, Z_i^x = L_1^x + L_2^x, L_3^x + L_4^x, \ldots, L_{2i-1}^x + L_{2i}^x$$

nimetame jada X blokipaaride pikkuste jadaks ning jada

$$Z^y(j) = Z_1^y, Z_2^y, \ldots, Z_j^y = L_1^y + L_2^y, L_3^y + L_4^y, \ldots, L_{2j-1}^y + L_{2j}^x$$

nimetame jada Y blokipaaride pikkuste jadaks.

Jada $\xi = \xi_1, \xi_2, \ldots, \xi_{i \wedge j}$ nimetame jadade X ja Y blokipaaride skooride jadaks, kus iga $k \in \{1, \ldots, i \wedge j\}$ korral

$$\xi_k = \begin{cases} 0, & kui\ X_1 \neq Y_1, \\ L_{2k-1}^x \wedge L_{2k-1}^y + L_{2k}^x \wedge L_{2k}^y, & kui\ X_1 = Y_1. \end{cases}$$

Näide. Kasutame jälle jadasid $x = 0011101111000$ ja $y = 10000110011$. Jada x blokipaaride pikkuste jada elemendid on $Z_1^x = 2+3 = 5, Z_2^x = 1+4 = 5$ ning jada y puhul $Z_1^y = 1+4 = 5, Z_2^y = 2+2 = 4$. Jadade x ja y blokipaaride skooride jada on praegusel juhul $\xi_1 = 0 = \xi_2$, aga kui jätaksime jada Y esimese bloki ära, siis saaksime $\xi_1 = 2 \wedge 4 + 3 \wedge 2 = 4$ ja $\xi_2 = 1 \wedge 2 + 4 \wedge 2 = 3$.

Jadade X ja Y bloki pikkuste jadad L^x, L^y, blokipaaride pikkuste jadad Z^x, Z^y ja nende blokipaaride skooride jada ξ on vahendid, mille abil saame tõestada tulemused vaadeldavate meetodite keskmiste skooride koondumise kohta, kuid enne seda tuleb näidata, et nendel vahenditel on kindlad omadused.

Teoreem 1. *Olgu meil lõpmatu, Bernoulli jaotusega i.i.d. jada $X = X_1 X_2 \ldots$. Siis jada L^x juhuslikud suurused on tinglikult sõltumatud tingimusel $X_1 = 0$ ja ka tingimusel $X_1 = 1$ ning jada L^x paarisarvulised liikmed on sama jaotusega ning ka paarituarvulised liikmed on sama jaotusega.*

Tõestus. Kõigepealt näitame, et jada L^x juhuslikud suurused on tinglikult sõltumatud tingimusel $X_1 = 0$. Analoogne tõestus käib juhu $X_1 = 1$ kohta. Definitsiooni kohaselt tahame näidata, et suvaliste naturaalarvude $n \geq 2, k_1, k_2, \ldots, k_n$ korral kehtib

$$P(L_1^x = k_1, L_2^x = k_2, \ldots, L_n^x = k_n | X_1 = 0) = \prod_{j=1}^{n} P(L_j^x = k_j | X_1 = 0).$$

Olgu $n, k_1, \ldots, k_n$ nõuetekohaselt, aga suvaliselt fikseeritud. Kuna me eeldame, et $X_1 = 0$, siis on iga $L_i^x, i \leq n$ puhul teada, mis värvi blokiga on tegu. Kui võtta ette suvaline blokk L_i^x, $i \leq n$, siis

kui i on paarisarv, siis $P(L_i^x = k_i | X_1 = 0) = (p^x)^{k_i - 1}(1 - p^x)$ ja

kui i on paaritu arv, siis $P(L_i^x = k_i | X_1 = 0) = (1 - p^x)^{k_i - 1}(p^x)$.

Paneme tähele, et iga L^x liige sõltub erinevatest jada X elementidest, seega, kui n on paarisarv, saame kirjutada, et

$$P(L_1^x = k_1, L_2^x = k_2, \ldots, L_n^x = k_n | X_1 = 0) =$$

$$= [(1-p^x)^{k_1-1}p^x][(p^x)^{k_2-1}(1-p^x)] \ldots [(p^x)^{k_n-1}(1-p^x)] = \prod_{j=1}^{n} P(L_j^x = k_j | X_1 = 0).$$

Kui n on paaritu, siis tuleb viimastesse kantsulgudesse korrutis $(1-p^x)^{k_n-1}(p^x)$. Seega jada L^x liikmed on tinglikult sõltumatud tingimusel $X_1 = 0$.

Kuna jada L^x paarituarvuliste indeksidega liikmed on 0- või 1-blokid ja paarisarvuliste indeksidega liikmed on vastavalt 1- või 0-blokid, siis saame näidata, et kõik paarituarvulised liikmed on sama jaotusega ja kõik paarisarvulised liikmed on sama jaotusega. Võtame kahe suvalise paarituarvulise bloki pikkused L_{2k-1}^x ja L_{2l-1}^x, kus $k, l \in \mathbb{N}$ ja $k \neq l$. Need on sama jaotusega, kui iga $m \in \mathbb{N}$ korral kehtib

$$P(L_{2k-1}^x = m) = P(L_{2l-1}^x = m).$$

Kuna L_{2k-1}^x loetleb mitu järjestikust tähte v esines vastavas blokis enne järgmise bloki algust ehk enne esimest $|v-1|$-tähte ning L_{2l-1}^x loetleb sama asja, aga erineva bloki kohta, siis

$$P(L_{2k-1}^x = m) = \sum_{j=\{0,1\}} P(L_{2k-1}^x = m | X_1 = j) P(X_1 = j) =$$

$$= \sum_{j=\{0,1\}} P(L_{2l-1}^x = m | X_1 = j) P(X_1 = j) = P(L_{2l-1}^x = m).$$

Seega jada L^x paarituarvuliste indeksidega liikmed on kõik sama jaotusega. Analoogne tõestus kehtib ka paarisarvuliste indeksidega liikmete kohta.

□

Teoreem 2. *Olgu meil lõpmatu, Bernoulli jaotusega i.i.d. jada $X = X_1X_2\ldots$. Siis jada X blokipaaride pikkuste jada Z^x on i.i.d. jada.*

Tõestus. Teoreem 1 ütleb, et jada L^x liikmed on tinglikult sõltumatud tingimusel $X_1 = 0$ ja tingimusel $X_1 = 1$. Näitame, et jada Z^x liikmed on sõltumatud. Kuna juhuslikud suurused $Z_1^x, Z_2^x, \ldots$ koosnevad kõik nii ühest 0-blokist kui ka ühest 1-blokist, siis iga $j, k \in \mathbb{N}$ korral

$$P(Z_j^x = k) = P(Z_j^x = k | X_1 = 0) = P(Z_j^x = k | X_1 = 1) \qquad (*)$$

ehk ei ole vahet, kas me teame esimese bloki värvi ning kui teame, siis ei ole vahet, kas see on 0 või 1. Fikseerime $n \geq 2, k_1, \ldots, k_n$ suvaliselt ning näitame täistõenäosuse valemi abil, et $P(Z_1^x = k_1, \ldots, Z_n^x = k_n) = \prod_{j=1}^n P(Z_j^x = k_j)$:

$$P(Z_1^x = k_1, \ldots, Z_n^x = k_n) =$$

$$\sum_{j\in\{0,1\}} P(Z_1^x = k_1, \ldots, Z_n^x = k_n | X_1 = j)P(X_1 = j) = \sum_{j\in\{0,1\}} P(X_1 = j)$$

$$\sum_{l_1^1+l_2^1=k_1} \ldots \sum_{l_1^n+l_2^n=k_n} P(L_1^x = l_1^1, L_2^x = l_2^1, \ldots, L_{2n-1}^x = l_1^n, L_{2n}^x = l_2^n | X_1 = j) =$$

$$= \sum_{j\in\{0,1\}} P(X_1 = j) \sum_{l_1^1+l_2^1=k_1} \ldots \sum_{l_1^n+l_2^n=k_n} P(L_1^x = l_1^1, L_2^x = l_2^1 | X_1 = j) \ldots$$

$$P(L_{2n-1}^x = l_1^n, L_{2n}^x = l_2^n | X_1 = j) =$$

$$= \sum_{j\in\{0,1\}} P(X_1 = j)P(Z_1^x = k_1 | X_1 = j) \ldots P(Z_n^x = k_n | X_1 = j) \overset{(*)}{=}$$

$$= \sum_{j\in\{0,1\}} P(X_1 = j)P(Z_1^x = k_1) \ldots P(Z_n^x = k_n) =$$

$$= P(Z_1^x = k_1) \ldots P(Z_n^x = k_n) \sum_{j\in\{0,1\}} P(X_1 = j) =$$

$$= P(Z_1^x = k_1) \ldots P(Z_n^x = k_n).$$

Nüüd näitame, et Z^x liikmed on kõik sama jaotusega ehk iga naturaalarvu $i, j, k \geq 2, i \neq j$ korral kehtib $P(Z_i^x = k) = P(Z_j^x = k)$, teades, et L^x paarisarvulised liikmed on sama jaotusega ning ka paarituarvulised liikmed on sama jaotusega. Fikseerime i, j, k suvaliselt nii, et $i \neq j$ ja $k \geq 2$. Teame tänu L^x liikmete tinglikule sõltumatusele, et

$$P(Z_i^x = k) = \sum_{m=\{0,1\}} \sum_{l_{2i-1}^x+l_{2i}^x=k} P(L_{2i-1}^x = l_{2i-1}^x, L_{2i}^x = l_{2i}^x | X_1 = m)P(X_1 = m) =$$

$$= \sum_{m=\{0,1\}} P(X_1 = m) \sum_{l_{2i-1}^x+l_{2i}^x=k} P(L_{2i-1}^x = l_{2i-1}^x | X_1 = m)P(L_{2i}^x = l_{2i}^x | X_1 = m)$$

ning analoogselt

$$P(Z_j^x = k) =$$

$$= \sum_{m=\{0,1\}} P(X_1 = m) \sum_{l_{2j-1}^x+l_{2j}^x=k} P(L_{2j-1}^x = l_{2j-1}^x | X_1 = m)P(L_{2j}^x = l_{2j}^x | X_1 = m).$$

Kuna nii Z_i^x kui ka Z_j^x puhul summeeritakse üle kahe muutuja, mille summa peab olema k, siis saame nende kahe summa liidetavatest moodustada paarid

$$P(L_{2i-1}^x = l_{2i-1}^x | X_1 = m) P(L_{2i}^x = l_{2i}^x | X_1 = m),$$
$$P(L_{2j-1}^x = l_{2j-1}^x | X_1 = m) P(L_{2j}^x = l_{2j}^x | X_1 = m),$$

kus $l_{2i-1}^x = l_{2j-1}^x$ ja $l_{2i}^x = l_{2j}^x$. Kuna jada L^x paarituarvulised liikmed on sama jaotusega ning paarisarvulised liikmed on sama jaotusega, siis iga paari puhul kehtib

$$P(L_{2i-1}^x = l_{2i-1}^x | X_1 = m) = P(L_{2j-1}^x = l_{2j-1}^x | X_1 = m) \text{ ja}$$
$$P(L_{2i}^x = l_{2i}^x | X_1 = m) = P(L_{2j}^x = l_{2j}^x | X_1 = m)$$

ning järelikult ka

$$P(L_{2i-1}^x = l_{2i-1}^x | X_1 = m) P(L_{2i}^x = l_{2i}^x | X_1 = m) =$$
$$= P(L_{2j-1}^x = l_{2j-1}^x | X_1 = m) P(L_{2j}^x = l_{2j}^x | X_1 = m)$$

ning saame, et

$$P(Z_i^x = k) = \sum_{m=\{0,1\}} P(X_1 = m)$$

$$\sum_{l_{2i-1}^x + l_{2i}^x = k} P(L_{2i-1}^x = l_{2i-1}^x | X_1 = m) P(L_{2i}^x = l_{2i}^x | X_1 = m) = \sum_{m=\{0,1\}} P(X_1 = m)$$

$$\sum_{l_{2j-1}^x + l_{2j}^x = k} P(L_{2j-1}^x = l_{2j-1}^x | X_1 = m) P(L_{2j}^x = l_{2j}^x | X_1 = m) = P(Z_j^x = k).$$

Seega juhuslike suuruste jada $Z_1^x, Z_2^x, \ldots$ on *i.i.d.* jada.

□

Teoreem 3. *Olgu meil lõpmatud, Bernoulli jaotustega, omavahel sõltumatud i.i.d. juhuslike suuruste jadad X ja Y. Siis nende blokipaaride skooride jada $\xi = \xi_1, \xi_2, \ldots$ on i.i.d. jada tingimusel $X_1 = Y_1$, s.t iga $n, k_1, \ldots, k_n \in \mathbb{N}$ puhul*

$$P(\xi_1 = k_1, \ldots, \xi_n = k_n | X_1 = Y_1) =$$
$$= P(\xi_1 = k_1 | X_1 = Y_1) P(\xi_1 = k_2 | X_1 = Y_1) \ldots P(\xi_1 = k_n | X_1 = Y_1).$$

Tõestus. Näitame, et ξ rahuldab seda tingimust. Selleks fikseerime naturaalarvu n ning defineerime jadade X ja Y blokipikkuste jadade põhjal juhuslikud vektorid

$$U_1 = (L_1^x, L_2^x, L_1^y, L_2^y), U_2 = (L_3^x, L_4^x, L_3^y, L_4^y), \ldots, U_n = (L_{2n-1}^x, L_{2n}^x, L_{2n-1}^y, L_{2n}^y).$$

Olgu $u_1 = (l_1^x, l_2^x, l_1^y, l_2^y), \ldots, u_n = (l_{2n-1}^x, l_{2n}^x, l_{2n-1}^y, l_{2n}^y) \in \mathbb{N}^4$ suvalised. Näitame, et vektorid $U_1, \ldots, U_n$ on *i.i.d.* tingimusel $X_1 = Y_1 = 0$. Teoreem 1 väidab, et jada L^x liikmed on sõltumatud tingimusel $X_1 = 0$ ning ka jada L^y liikmed on sõltumatud tingimusel $Y_1 = 0$. Kuna jadad X ja Y on omavahel sõltumatud, siis saame, et ka jadad L^x ja L^y on omavahel sõltumatud. Kirjutame vektorid $U_i, i \in \mathbb{N}^4$ lahti ja kasutame tingliku ja tingimatu sõltumatuse omadusi:

$$P(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n | X_1 = Y_1 = 0) =$$

$$= P(L_1^x = l_1^x, L_2^x = l_2^x, L_1^y = l_1^y, L_2^y = l_2^y, L_3^x = l_3^x, \ldots, L_{2n}^y = l_{2n}^y | X_1 = Y_1 = 0) =$$

$$= P(L_1^x = l_1^x, L_2^x = l_2^x, L_1^y = l_1^y, L_2^y = l_2^y | X_1 = Y_1 = 0)$$

$$P(L_3^x = l_3^x, L_4^x = l_4^x, L_3^y = l_3^y, L_4^y = l_4^y | X_1 = Y_1 = 0) \ldots$$

$$P(L_{2n-1}^x = l_{2n-1}^x, L_{2n}^x = l_{2n}^x, L_{2n-1}^y = l_{2n-1}^y, L_{2n}^y = l_{2n}^y | X_1 = Y_1 = 0) = \ldots$$

Teoreem 1 väidab, et jada L^x paarisarvulised liikmed on sama jaotusega ning ütleb sama jada L^x paarituarvuliste liikmete ning jada L^y paaris- ja paarituarvuliste liikmete kohta. Kasutades neid teadmisi ja jälle sõltumatuse omadusi, saame kirjutada

$$= P(L_1^x = l_1^x, L_2^x = l_2^x, L_1^y = l_1^y, L_2^y = l_2^y | X_1 = Y_1 = 0)$$

$$P(L_1^x = l_3^x, L_2^x = l_4^x, L_1^y = l_3^y, L_2^y = l_4^y | X_1 = Y_1 = 0) \ldots$$

$$P(L_1^x = l_{2n-1}^x, L_2^x = l_{2n}^x, L_1^y = l_{2n-1}^y, L_2^y = l_{2n}^y | X_1 = Y_1 = 0) =$$

$$= P(U_1 = u_1 | X_1 = Y_1 = 0) P(U_1 = u_2 | X_1 = Y_1 = 0) \ldots P(U_1 = u_n | X_1 = Y_1 = 0).$$

Sama arutelu kehtib tingimusel $X_1 = Y_1 = 1$, kus saame suvaliste $u_1, \ldots, u_n \in \mathbb{N}^4$ korral, et

$$P(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n | X_1 = Y_1 = 1) =$$

$$= P(U_1 = u_1 | X_1 = Y_1 = 1) \ldots P(U_1 = u_n | X_1 = Y_1 = 1).$$

Olgu $f : \mathbb{N}^4 \to \mathbb{N}$ defineeritud kui $f((a, b, c, d)) = a \wedge c + b \wedge d$. Seega $f(U_i) = L_{2i-1}^x \wedge L_{2i-1}^y + L_{2i}^x \wedge L_{2i}^y = \xi_i, i = 1, 2, \ldots$. Näitame, et ka jada $\xi_1, \ldots, \xi_n$ on tingimusel $X_1 = Y_1 = 0$ *i.i.d.*. Fikseerime $k_1, \ldots, k_n \in \mathbb{N}$ suvaliselt ja kasutame eelnevat *i.i.d.* jadaks olemise tingimust, teades, et jada $U_1, \ldots, U_n$ on tingimusel $X_1 = Y_1 = 0$ *i.i.d.* jada:

$$P(\xi_1 = k_1, \dots, \xi_n = k_n | X_1 = Y_1 = 0) =$$
$$P(f(U_1) = k_1, \dots, f(U_n) = k_n | X_1 = Y_1 = 0) =$$
$$= P(U_1 \in f^{-1}(k_1), \dots, U_n \in f^{-1}(k_n) | X_1 = Y_1 = 0) =$$
$$= P(U_1 \in f^{-1}(k_1) | X_1 = Y_1 = 0) \dots P(U_n \in f^{-1}(k_n) | X_1 = Y_1 = 0) =$$
$$= P(U_1 \in f^{-1}(k_1) | X_1 = Y_1 = 0) \dots P(U_1 \in f^{-1}(k_n) | X_1 = Y_1 = 0) =$$
$$= P(f(U_1) = k_1 | X_1 = Y_1 = 0) \dots P(f(U_1) = k_n | X_1 = Y_1 = 0)$$
$$= P(\xi_1 = k_1 | X_1 = Y_1 = 0) \dots P(\xi_1 = k_n | X_1 = Y_1 = 0).$$

Analoogselt saab näidata, et jada $\xi_1, \dots, \xi_n$ on *i.i.d.* tingimusel $X_1 = Y_1 = 1$ ehk

$$P(\xi_1 = k_1, \dots, \xi_n = k_n | X_1 = Y_1 = 1) =$$
$$= P(\xi_1 = k_1 | X_1 = Y_1 = 1) \dots P(\xi_1 = k_n | X_1 = Y_1 = 1).$$

Paneme tähele, et juhusliku suuruse $\xi_i, i \in \mathbb{N}$ jaotus ei sõltu sellest, kas me teame, milline on jada X esimene liige või mitte ning kui teame, ei ole vahet, kas esimene liige on 0 või 1. S.t iga $i \in \{1, \dots, n\}$ ja suvalise $k \in \mathbb{N}$ korral

$$P(\xi_i = k | X_1 = Y_1 = 0) = P(\xi_i = k | X_1 = Y_1 = 1) = P(\xi_i = k | X_1 = Y_1).$$

Võrdused kehtivad, sest iga tingimuse korral sõltub ξ_i ühest 0-blokkide paarist ja ühest 1-blokkide paarist ning nende paaride skoorid liidetakse kokku, mis on sümmeetriline tehe. Kokkuvõttes saame, et suvaliste naturaalarvude $k_1, \dots, k_n$ korral

$$P(\xi_1 = k_1, \dots, \xi_n = k_n | X_1 = Y_1) =$$
$$= P(\xi_1 = k_1 | X_1 = Y_1) P(\xi_1 = k_2 | X_1 = Y_1) \dots P(\xi_1 = k_n | X_1 = Y_1).$$

□

Vormistame eraldi tulemuse jadade Z^x, Z^y ja ξ liikmete keskväärtuste leidmiseks.

Teoreem 4. *Olgu meil lõpmatud, omavahel sõltumatud, Bernoulli jaotustega i.i.d. jadad X ja Y, kusjuures $P(X_1 = 1) = p^x$ ja $P(Y_1 = 1) = p^y$ ning eeldame, et $X_1 = Y_1$. Siis jadade X ja Y blokipaaride pikkuste jadade ning nende blokipaaride skooride jada liikmete keskväärtused avalduvad järgnevalt:*

- $\mu^x := E(Z_1^x) = \frac{1}{p^x(1-p^x)}$ *ja* $\mu^y := E(Z_1^y) = \frac{1}{p^y(1-p^y)}$,
- $m := E(\xi_1 | X_1 = Y_1) = \frac{1}{p^x + p^y - p^x p^y} + \frac{1}{1 - p^x p^y}$.

Tõestus. Leiame väärtused μ^x, μ^y ja m. Kuna teoreemid 2 ja 3 väidavad, et jadad Z^x, Z^y ja ξ on *i.i.d.* jadad, siis piisab iga jada puhul leida vaid esimese liikme keskväärtus.

Ei ole teada, mis on jada X esimene värv, aga teame, et Z_1^x koosneb ühest 0-blokist ja ühest 1-blokist, seega tema keskväärtuse saame arvutada liites kokku ühe 0-bloki keskväärtuse ning ühe 1-bloki keskväärtuse. Kuna suvalise 0-bloki pikkus on geomeetrilise jaotusega juhuslik suurus parameetriga p^x ja suvaline 1-blokk vastavalt parameetriga $1-p^x$, siis

$$\mu^x = EZ_1^x = E(L_1^x|X_1=0) + E(L_2^x|X_1=0) = \frac{1}{1-p^x} + \frac{1}{p^x} = \frac{1}{p^x(1-p^x)}.$$

Analoogselt saame jada Y kohta, et

$$\mu^y = EZ_1^y = \frac{1}{p^y(1-p^y)}.$$

Juhusliku suuruse ξ_1 keskväärtuse m leiame sarnasel moel. Kuna me eeldame, et joondatavate jadade esimesed liikmed on ühte värvi, siis teame, et iga $i \in \mathbb{N}$ korral koosneb ξ_i ühest joondatud 0-blokkide paari skoorist ja ühest joondatud 1-blokkide paari skoorist, mistõttu

$$\begin{aligned} E(\xi_1|X_1=Y_1) &= E(L_1^x \wedge L_1^y + L_2^x \wedge L_2^y | X_1=Y_1) = \\ &= E(L_1^x \wedge L_1^y|X_1=0) + E(L_2^x \wedge L_2^y|X_1=0) = \\ &= E(L_1^x \wedge L_1^y|X_1=1) + E(L_2^x \wedge L_2^y|X_1=1). \end{aligned}$$

Näitame, et kui $X_1 = Y_1 = 0$, siis $L_1^x \wedge L_1^y \sim Geom(1-(1-p^x)(1-p^y))$. Selleks piisab näidata, et

$$P(L_1^x \wedge L_1^y > k | X_1 = Y_1 = 0) = (1-(1-(1-p^x)(1-p^y)))^k.$$

Kuna X ja Y on sõltumatud, siis

$$P(L_1^x \wedge L_1^y > k | X_1 = Y_1 = 0) = P(L_1^x > k|X_1=0)P(L_1^y > k|Y_1=0).$$

Kasutades jällegi teadmist, et L_1^x ja L_1^y on geomeetrilise jaotusega, saame, et

$$\begin{aligned} P(L_1^x > k|X_1=0)P(L_1^y > k|Y_1=0) &= (1-p^x)^k(1-p^y)^k = \\ &= ((1-p^x)(1-p^y))^k = (1-(1-(1-p^x)(1-p^y)))^k. \end{aligned}$$

Analoogselt saame, et kui $X_1 = Y_1 = 0$, siis $L_2^x \wedge L_2^y \sim Geom(1-p^xp^y)$. Kokkuvõttes

$$E(\xi_1|X_1=Y_1) = \frac{1}{1-(1-p^x)(1-p^y)} + \frac{1}{1-p^xp^y} = \frac{1}{p^x+p^y-p^xp^y} + \frac{1}{1-p^xp^y}.$$

□

Sellega on tõestatud vajalikud omadused meie abivahendite kohta ning nüüd tõestame üldisema tulemuse, mis aitab hiljem leida bloki kaupa Hammingu meetodi ja tähe kaupa järjestikuse meetodi keskmise skoori piirväärtuse.

Teoreem 5. *Olgu* $Z = Z_1, Z_2, \ldots$ *i.i.d. juhuslikud suurused, kusjuures* $\mathrm{P}(Z_1 \geq 1) = 1$ *ja* $EZ_1 = \mu < \infty$. *Defineerime*

$$T_0 = 0, \quad T_k = Z_1 + \ldots Z_k, \quad k = 1, 2, \ldots$$

ja

$$M(n) = \max\{k = 0, 1, 2, \ldots : T_k \leq n\}.$$

Kui $n \to \infty$, *siis kehtib koondumine*

$$\frac{M(n)}{n} \to \frac{1}{\mu}, \quad \text{p.k..}$$

Tõestus. Alustuseks näitame, et kui $n \to \infty$, siis $M(n) \to \infty$, p.k.. Näeme, et n kasvades kasvab ka $M(n)$, sest viimane on defineeritud kui suurim indeks k nii, et T_k on ülalt tõkestatud arvuga n. Ainus juhus, mille korral $M(n)$ on tõkestatud on see, et leidub selline Z_i, mille väärtus on lõpmatult suur. Sel juhul kehtiks piisavalt suure n korral $T_{i-1} \leq n$, aga $T_i > n$ ehk $M(n)$ oleks ülimalt i. Selline juhus esineb tõenäosusega 0, sest $EZ_1 = \mu < \infty$ ja $Z_1, Z_2, \ldots$ on *i.i.d.* juhuslikud suurused. Seega kehtib p.k. koondumine $M(n) \to \infty$.

Tugevast suurte arvude seadusest saame, et kui $k \to \infty$, siis

$$\frac{T_k}{k} \to \mu, \quad \text{p.k.}$$

ja, et $\lim\limits_{n\to\infty} M(n) = \infty$ p.k., siis kehtib ka koondumine

$$\frac{T_{M(n)}}{M(n)} \to \mu, \quad \text{p.k..}$$

Kuna $M(n)$ valib välja sellise indeksi k, et kehtiks $T_k \leq n$, siis teame, et $T_{M(n)} \leq n$ ning kuna $M(n)$ valib välja suurima indeksi, mis seda tingimust rahuldab, siis teame, et $n < T_{M(n)+1}$. Neid võrratusi muutujaga $M(n)$ läbi jagades saame

$$\frac{T_{M(n)}}{M(n)} \leq \frac{n}{M(n)} < \frac{T_{M(n)+1}}{M(n)}.$$

Kuna $\frac{T_{M(n)}}{M(n)} \to \mu$, p.k. ja koondumise $\lim\limits_{n\to\infty} \frac{M(n)+1}{M(n)} = 1$, p.k., abil teame, et

$$\lim_{n\to\infty}\frac{T_{M(n)+1}}{M(n)}=\lim_{n\to\infty}\left(\frac{T_{M(n)+1}}{M(n)+1}\frac{M(n)+1}{M(n)}\right)=$$

$$=\lim_{n\to\infty}\left(\frac{T_{M(n)+1}}{M(n)+1}\right)\cdot 1=\lim_{n\to\infty}\left(\frac{T_{M(n)}}{M(n)}\right)=\mu,\quad \text{p.k.},$$

siis võileivateoreemi abil teame, et kehtib ka $\frac{n}{M(n)}\to\mu,$ p.k., millest järeldub

$$\frac{M(n)}{n}\to\frac{1}{\mu},\quad \text{p.k.}.$$

□

1.3 Joonduste skoorid

Selles peatükis vaatleme kolme joondust, mida on uuritud magistritöös Toots (2012). Nendeks on Hammingu joondus, bloki kaupa Hammingu joondus ja tähe kaupa järjestikune joondus. Iga joonduse puhul defineerime selle skoori ning tõestame tulemuse selle keskmise skoori piirväärtuse leidmiseks, mis on saadud Tootsi magistritöö eeskujul.

1.3.1 Hammingu joondus

Joondused, mida selles töös uurime, on naiivsed ja mitte eriti rafineeritud. Kuid neist kõige naiivsem on Hammingu joondus, sest aktiivset joondamist ei toimugi - võrreldavad jadad asetatakse üksteise kohale ning loetakse kokku *match*'id ja *mismatch*'id. Tootsi magistritöös vaadeldi sellist Hammingu joondust, kus loeti kokku ainult *match*'id, kuid ka *mismatch*'ide arvestamine annab parema arusaama jadade sarnasusest. Hammingu joondus on huvitav mõõdupuu teistele joondustele, sest kui mõne teise joonduse skoor ei ole palju parem Hammingu joonduse omast või see annab lausa halvema skoori, siis võime öelda, et tegu ei ole hea joondusega. Teisisõnu moodustab Hammingu joondus alumise piiri kõigile joondustele, sest Hammingu joondus annab selle skoori, mille saaksime ilma igasuguse pingutuseta.

Definitsioon 16. *Olgu $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$ lõpmatud jadad. Jadade X ja Y Hammingu joonduseks nimetatakse nende joondust ehk paari (X, Y). Hammingu joonduse skoori saame leida võrduse*

$$B_n=\sum_{i=1}^{n}\mathbb{1}_{\{X_i=Y_i\}}$$

abil, kus

$$\mathbb{1}_{\{X_i=Y_i\}} = \begin{cases} 1, & kui\ X_i = Y_i \\ -1, & kui\ X_i \neq Y_i \end{cases}$$

iga $i \in 1, \ldots, n$.

Teoreem 6. *Olgu* $X = X_1X_2\ldots$ *ja* $Y = Y_1Y_2\ldots$ *lõpmatud, omavahel sõltumatud, Bernoulli jaotustega i.i.d. jadad, kusjuures* $P(X_1 = 1) = p^x$ *ja* $P(Y_1 = 1) = p^y$. *Siis jadade* X *ja* Y *Hammingu joonduse skoori* B_n *kohta kehtib järgnev p.k. koondumine*

$$\frac{B_n}{n} \to \gamma_H, \ p.k.$$

ning keskmise skoori piirväärtus on

$$\gamma_H := (2p^x - 1)(2p^y - 1).$$

Tõestus. Leiame Hammingu joonduse korral ühe joondatud elementide paari skoori keskväärtuse, teades, et X ja Y on omavahel sõltumatud:

$$\begin{aligned}
&\mathrm{E}(\mathbb{1}_{\{X_i=Y_i\}}) = \\
&= 1 \cdot p^x p^y + 1 \cdot (1-p^x)(1-p^y) + (-1) \cdot p^x(1-p^y) + (-1) \cdot (1-p^x)p^y = \\
&= p^x p^y + (1-p^x)(1-p^y) - (p^x(1-p^y) + (1-p^x)p^y) = \\
&= p^x p^y + 1 - p^y - p^x + p^x p^y - (p^x - p^x p^y + p^y - p^x p^y) = \\
&= 4p^x p^y - 2p^x - 2p^y + 1 = 2p^x(2p^y - 1) - (2p^y - 1) = \\
&= (2p^x - 1)(2p^y - 1).
\end{aligned}$$

Kuna X ja Y on *i.i.d.* jadad, X ja Y on omavahel sõltumatud ja $E(\mathbb{1}_{\{X_i=Y_i\}}) < \infty$, siis tugeva suurte arvude seaduse abil teame, et

$$\frac{B_n}{n} = \frac{\sum_{i=1}^n \mathbb{1}_{\{X_i=Y_i\}}}{n} \to (2p^x - 1)(2p^y - 1) =: \gamma_H, \ \text{p.k.}$$

□

1.3.2 Bloki kaupa Hammingu joondus

Hammingu joonduse ideed saab rakendada ka jadadele, mis on esitatud blokkidena. Kuna kahetähelise tähestiku puhul on mittetriviaalne juhtum vaid see, kus esimesed joondatud blokid on sama värvi, siis vaatleme sellist bloki kaupa Hammingu joondust, kus teisel jadal jäetakse vajadusel esimene blokk algusest ära. Edasi asetame kummagi jada k-ndad blokid vastavusse, lühemale nendest lisame lõppu nii

palju *indel*'eid, et saadud blokid oleksid võrdse pikkusega ning skoori saamiseks loeme kokku blokipaaride lühema bloki pikkuse. Kui jadadel on erinev arv blokke, siis joondame kuni väiksema blokkide arvuga jada viimase tervikliku blokini ja ülejäänud elemendid selles jadas joondame teise jada järgmise blokiga ehk väiksema blokkide arvuga jada kasutame nö tervikuna ära. Järgnevalt võtame kasutusele tähised, et defineerida bloki kaupa Hammingu joondus.

Olgu $S_k^x := L_1^x + L_2^x + \ldots + L_k^x$, $S_k^y := L_1^y + L_2^y + \ldots + L_k^y$ ning $K^x(n) := \max\{k \in \mathbb{N} : S_k^x < n\}$ ja $K^y(n) := \max\{k \in \mathbb{N} : S_k^y < n\}$. Ehk $K^x(n)$ tähistab jada X täisblokkide arvu kuni indeksini n. Olgu

$$M^x(n) := \begin{cases} \frac{K^x(n)}{2}, & \text{kui } K^x(n) \text{ on paarisarv,} \\ \frac{K^x(n)-1}{2}, & \text{kui } K^x(n) \text{ on paaritu arv} \end{cases}$$

ehk $M^x(n)$ tähistab jada X täisblokipaaride arvu ja olgu $M^y(n)$ defineeritud $K^y(n)$ põhjal analoogselt. Olgu $L^x_{poolik}(n) := n - S_{K^x(n)}$ ja $L^y_{poolik}(n) := n - S_{K^y(n)}$, mis tähistavad vastavate jadade viimaste ehk lõpetamata blokkide pikkust.

Näide. Olgu $x = 000111100111100111$ ja $n = 11$. Siis $K^x(11) = 3$, $M^x(11) = 1$ ja $L^x_{poolik}(11) = 2$. Kui $n = 13$, siis ikkagi $K^x(13) = 3$, aga $L^x_{poolik}(13) = 4$.

Olgu $\tau := \min\{k \in \mathbb{N} : Y_k = X_1\}$ ja defineerime uue jada $Y' = Y'_1, Y'_2, \ldots$ selliselt, et $Y'_k = Y_{\tau+k-1}$, $k \in \mathbb{N}$ ehk Y' on jada Y, millel jätsime vajadusel esimese bloki ära. Tähistagu $L' = L'_1, L'_2, \ldots$ jada Y' bloki pikkuste jada, $K'(n)$ selle täisblokkide arvu indeksini n, $M'(n)$ selle täisblokipaaride arvu indeksini n, $L'_{poolik}(n)$ selle lõpetamata bloki pikkust indeksini n ja $\xi' = \xi'_1, \xi'_2, \ldots$ olgu jadade X ja Y' blokipaaride skooride jada.

Näide. Kasutame eelmise näite jada $x = 000111100111100111$ ning jada $y = 110011000011000011$. Nende jadade puhul on $\tau = 3$ ning saame uue jada $y' = 0011000011000011$, mille puhul $K'(13) = 4, M'(13) = 2, L'_{poolik} = 1$. Jadade x ja y' blokipaaride skooridest on defineeritud ainult esimene ja $\xi'_1 = 3 \wedge 2 + 4 \wedge 2 = 4$.

Definitsioon 17. *Vaatleme jadade $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$ esimest n elementi. Kui $\tau > n$, siis nende bloki kaupa Hammingu joonduse skoor on $B^H(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = 0$, sest jadas Y ei leidu ühtegi täisblokki, mida joondada. Kui aga $\tau \leq n$, siis skooriks saame*

$$B^H(X_1, \ldots, X_n, Y_1, \ldots, Y_n) :=$$
$$:= B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_{n-(\tau-1)}) = \sum_{i=1}^{\mathcal{K}(n)} L_i^x \wedge L'_i + V(n),$$

kus $\mathcal{K}(n) := K^x(n) \wedge K'(n - (\tau - 1))$ ja

$$V(n) := \begin{cases} L^x_{poolik}(n) \wedge L'_{poolik}(n), & kui\ K^x(n) = K'(n-(\tau-1)), \\ L^x_{poolik}(n) \wedge L'_{\mathcal{K}(n)+1}, & kui\ K^x(n) < K'(n-(\tau-1)), \\ L^x_{\mathcal{K}(n)+1} \wedge L'_{poolik}(n), & kui\ K'(n-(\tau-1)) < K^x(n). \end{cases}$$

Näide. Kasutame eelmise näite jadasid

$$x = 000111100111100111$$
$$y' = 0011000011000011$$

ja olgu $n = 13$. Näeme, et $\mathcal{K}(13) = K^x(13) \wedge K'(13-(3-1)) = 3 \wedge 4 = 3$ ning kuna praegusel juhul $K^x(13) < K'(11)$, siis $V(13) = L^x_{poolik}(13) \wedge L'_{\mathcal{K}(13)+1} = L^x_{poolik}(13) \wedge L'_4 = 4 \wedge 2 = 2$. Nüüd saame arvutada skoori, milleks tuleb

$$B^H(000111100111100111, 0011000011000011) = 3 \wedge 2 + 4 \wedge 2 + 2 \wedge 4 + 2 = 8.$$

Tõestame abitulemuse, mida läheb vaja järgmises lemmas.

Lemma 1. *Olgu antud i.i.d. jada $A_1, A_2, \ldots$, kusjuures $E(A_1) < \infty$. Siis piirprotsessis $n \to \infty$ kehtib koondumine*

$$\frac{A_n}{n} \to 0, p.k..$$

Tõestus. Kasutades eeldust, et $E(A_1) < \infty$ ja tugevat suurte arvude seadust saame, et

$$\frac{A_n}{n} = \frac{(A_1 + A_2 + \ldots + A_n) - (A_1 + A_2 + \ldots + A_{n-1})}{n} =$$
$$= \frac{A_1 + A_2 + \ldots + A_n}{n} - \frac{n-1}{n} \cdot \frac{A_1 + A_2 + \ldots + A_{n-1}}{n-1} \to$$
$$\to E(A_1) - 1 \cdot E(A_1) = 0, p.k..$$

□

Me tahame teada, milliseks konstandiks koondub $\frac{B^H(X(n),Y'(n-(\tau-1))}{n}$ piirprotsessis $n \to \infty$ ning selleks me näitame algul järgmises lemmas, et meid huvitav koondumine on ekvivalentne ühe lihtsama koondumisega ja seejärel tõestame tulemuse selle lihtsama koondumise kohta.

Lemma 2. *Olgu meil lõpmatud, omavahel sõltumatud, Bernoulli jaotusega i.i.d. juhuslike suuruste jadad X ja Y. Siis piirprotsessis $n \to \infty$ on koondumine*

$$\frac{B^H_n}{n} := \frac{B^H(X_1, \ldots, X_n, Y_1, \ldots, Y_n)}{n} \to \gamma_{BH}, p.k., \tag{1}$$

ekvivalentne koondumisega

$$\frac{B_n^B}{n} := \frac{B^B(X_1,\dots,X_n,Y_1',\dots,Y_n')}{n} \to \gamma_{BH}, p.k., \tag{2}$$

kus $B^B(X_1,\dots,X_n,Y_1',\dots,Y_n') := \sum_{i=1}^{M^x(n)\wedge M'(n)} \xi_i'$ *ja* γ_{BH} *on reaalarvuline konstant.*

Tõestus. Näitame kõigepealt, et koondumine (1) on ekvivalentne koondumisega

$$\frac{B^H(X_1,\dots,X_n,Y_1,\dots,Y_{n+\tau-1})}{n} = \frac{B^H(X_1,\dots,X_n,Y_1',\dots,Y_n')}{n} \to \gamma_{BH}, p.k.. \tag{3}$$

Kuna koondumises (3) kasutatakse skoori leidmiseks $\tau-1$ rohkem elementi, siis võib $\tau \geq 2$ korral olla see skoor suurem, kuid paneme tähele, et skoor saab suureneda vaid $\tau - 1$ võrra. Kuna uued elemendid tekivad jada $Y_1,\dots,Y_n$ lõppu juurde, aga joondamine algab jada algusest, siis teame, et esimesed n elementi joondatakse ka uute elementide lisamisel endist viisi. Ja kuna elemente lisatakse juurde $\tau-1$ tükki, siis ei saa skoor suureneda rohkem kui $\tau - 1$ võrra. Järelikult

$$0 \leq B^H(X_1,\dots,X_n,Y_1',\dots,Y_n') - B^H(X_1,\dots,X_n,Y_1,\dots,Y_n) \leq \tau - 1.$$

Jagades läbi muutujaga n saame, et piirprotsessis $n \to \infty$ kehtib

$$0 \leq \frac{B^H(X_1,\dots,X_n,Y_1',\dots,Y_n')}{n} - \frac{B^H(X_1,\dots,X_n,Y_1,\dots,Y_n)}{n} \leq \frac{\tau-1}{n}, p.k..$$

Kuna $\tau - 1$ on arv, siis teame, et kehtib p.k. koondumine

$$\frac{\tau-1}{n} \to 0, p.k..$$

Järelikult on koondumised (1) ja (3) ekvivalentsed.

Näitame nüüd, et ka koondumised (3) ja (2) on ekvivalentsed. Olgu

$$R^x(n) := n - \sum_{i=1}^{M^x(n)} Z_i \text{ ja } R'(n) := n - \sum_{i=1}^{M'(n)} Z_i',$$

kus jada $Z' = Z_1', Z_2', \dots$ on defineeritud jada L' põhjal. Kehtivad võrratused

$$\sum_{i=1}^{M^x(n)} Z_i < n \leq \sum_{i=1}^{M^x(n)+1} Z_i,$$

millest saame

$$0 < R^x(n) \leq Z^x_{M^x(n)+1}.$$

Paneme tähele, et

$$B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) < B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n),$$

sest B^H skoor kujuneb alguses täpselt samamoodi nagu B^B skoor kuni blokkideni $Z^x_{M(n)}$ ja $Z'_{M(n)}$, $M(n) := M^x(n) \wedge M'(n)$. Aga sealt edasi lisatakse B^H skoorile võimalusel juurde ühe täisblokipaari skoor, kui $M(n)$ on paaritu arv ning kindlasti lisatakse juurde $V(n)$. Näitame, et

$$B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) \leq B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) + R^x(n) \vee R'(n).$$

Olgu üldisust kitsendamata $M(n) = M^x(n)$. Suurus $R^x(n)$ tähistab seda, mitu elementi jääb üle kuni indeksini n, kui oleme kõik jada X täisblokipaarid ära joondanud ehk mitu elementi saaksime veel kasutada skoori B^H arvutamiseks, kui skoori B^B osa on arvutatud. Siit näemegi, et skoor B^H ei saa olla skoorist B^B suurem rohkem kui $R^x(n)$ võrra, sest täpselt nii palju on meil elemente, mida kasutada skoori arvutamiseks pärast täisblokipaaride skooride arvutamist. Juhul $M(n) = M'(n)$ on ülemiseks tõkkeks $R'(n)$, mistõttu saame üldjuhul skooride vahet hinnata nende kahe maksimumiga. Seega teame, et

$$B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) < B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) \leq$$
$$\leq B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) + R^x(n) \vee R'(n),$$

millest B^B skoori lahutades saame

$$0 < B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) - B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) \leq$$
$$R^x(n) \vee R'(n) \leq Z^x_{M^x(n)+1} \vee Z'_{M'(n)+1} \leq Z^x_{M^x(n)+1} + Z'_{M'(n)+1},$$

millest omakorda

$$\frac{B^H(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n) - B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n)}{n} \leq$$
$$\leq \frac{Z^x_{M^x(n)+1} + Z'_{M'(n)+1}}{n} = \frac{Z^x_{M^x(n)+1}}{n} + \frac{Z'_{M'(n)+1}}{n}.$$

Teoreemis 5 nägime analoogses olukorras, et kui $n \to \infty$, siis ka $M^x(n) \to \infty$ ja $M'(n) \to \infty$, mistõttu saame lemma 1 abil, et

$$\frac{Z^x_{M^x(n)+1}}{n} + \frac{Z'_{M'(n)+1}}{n} \to 0 + 0 = 0, p.k..$$

Järelikult on koondumine (3) ekvivalentne koondumisega (2).

□

Nüüd saame näidata, et skoori B_n^B keskmine koondub konstandiks γ_{BH} ning lemma 2 põhjal teame, et siis koondub ka skoori B_n^H keskmine samaks konstandiks.

Teoreem 7. *Olgu $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$ lõpmatud, omavahel sõltumatud, Bernoulli jaotustega i.i.d. jadad, kusjuures $P(X_1 = 1) = p^x$ ja $P(Y_1 = 1) = p^y$. Siis kehtib p.k. koondumine*

$$\frac{B^B(X_1,\ldots,X_n,Y_1',\ldots,Y_n')}{n} = \frac{\sum_{i=1}^{M^x(n)\wedge M'(n)}\xi_i'}{n} \to \Big(\frac{1}{\mu^x}\wedge\frac{1}{\mu^y}\Big)m =: \gamma_{BH},$$

kus $\mu^x = E(Z_1^x) = \frac{1}{p^x(1-p^x)}, \mu^y = E(Z_1^y) = \frac{1}{p^y(1-p^y)}$ ning $m = E(\xi_1) = \frac{1}{p^x+p^y-p^xp^y}$ $+ \frac{1}{1-p^xp^y}$.

Tõestus. Siin kasutame jada $Y = Y_1, Y_2, \ldots$ asemel jada $Y' = Y_1', Y_2', \ldots$. Kuna $Y_1' = X_1$, siis on Y_1' sama jaotusega, mis X_1, kuid ikkagi $\forall i = 2, 3, \ldots, Y_i' \sim Y_2'$. Lisaks on jada Y' liikmed ikkagi omavahel sõltumatud. Meie jaoks on oluline, et ka jada Y' kohta kehtiksid teoreemide 1, 2 ja 3 tulemused. Esimene neist kehtib selle pärast, et teoreemi 1 tõestuses ei ole oluline, et esimene liige Y_1' oleks sama jaotusega, mis kõik teised, sest L_1' arvutades me eeldame, et esimene liige on juba kindlaks määratud. Ehk ikkagi $L_1' \sim L_{2l-1}', l \in \mathbb{N}$.

Teoreem 2 tugineb teoreemile 1, seega selle väide kehtib ka jada Y' kohta. Teoreemis 3 eeldame, et $X_1 = Y_1$, aga kui rakendada seda teoreemi jadadele X ja Y', siis kehtib $X_1 = Y_1'$ ka ilma eelduseta. Teisisõnu ei ole juhuslike suuruste X_1 ja Y_1' sõltuvus teoreemis 3 piiranguks ning tänu sellele sõltuvusele saame, et jada $\xi' = \xi_1', \xi_2', \ldots$ on tingimatult *i.i.d.*.

Teoreemi 2 põhjal teame, et Z^x on *i.i.d.* jada. Lisaks teame, et $P(Z_1^x \geq 1) = 1$ ja teoreemi 4 põhjal, et $EZ_1^x < \infty$ ehk kokkuvõttes on teoreemi 5 eeldused täidetud. Defineerime jada $Z_1^x, Z_2^x, \ldots$ põhjal iga $k \in \mathbb{N}$ jaoks $T_k^x = \sum_{j=1}^{k} L_j^x$ ning $M^x(n) = \max\{k \in \mathbb{N} | T_k^x < n\}$ ja $\mu^x = E(Z_1^x)$. Saame, et

$$\frac{M^x(n)}{n} \to \frac{1}{\mu^x}, \quad \text{p.k..}$$

Samamoodi rahuldab teoreemi 5 tingimusi jada Z', seega saame analoogselt defineerida jada $Z_1', Z_2', \ldots$ põhjal iga $k \in \mathbb{N}$ jaoks T_k', $M'(n)$ ja μ' ning teame, et kehtib

$$\frac{M'(n)}{n} \to \frac{1}{\mu'}, \quad \text{p.k..}$$

Defineerime $M(n) = M^x(n) \wedge M'(n)$ ning saame

$$\frac{M(n)}{n} = \frac{M^x(n)}{n} \wedge \frac{M'(n)}{n} \to \frac{1}{\mu^x} \wedge \frac{1}{\mu'}, \quad \text{p.k..}$$

Kuna teoreemi 4 põhjal $m = E(\xi'_1) < \infty$ ja teoreemi 3 ja toodud selgituste põhjal on ξ' *i.i.d.* jada, siis tugeva suurte arvude seaduse abil teame, et kui $k \to \infty$, siis

$$\frac{1}{k} \sum_{i=1}^{k} \xi'_i \to m, \quad \text{p.k..}$$

Teoreemiga 5 analoogselt teame, et kui $n \to \infty$, siis kehtib ka koondumine

$$\frac{1}{M(n)} \sum_{i=1}^{M(n)} \xi'_i \to m, \quad \text{p.k..}$$

Vaatleme summat $\frac{1}{n} \sum_{i=1}^{M(n)} \xi'_i$ piirprotsessis $n \to \infty$ ning saame koondumise

$$\frac{B^B(X_1, \ldots, X_n, Y'_1, \ldots, Y'_n)}{n} = \frac{1}{n} \sum_{i=1}^{M(n)} \xi'_i =$$

$$= \frac{M(n)}{n} \cdot \frac{1}{M(n)} \sum_{i=1}^{M(n)} \xi'_i \to \left(\frac{1}{\mu^x} \wedge \frac{1}{\mu'}\right) \cdot m = \gamma_{BH}, \quad \text{p.k..}$$

Teoreemi 4 põhjal teame, et

$$\gamma_{BH} = \left(\frac{1}{\mu^x} \wedge \frac{1}{\mu'}\right) \cdot m = \left(p^x(1-p^x) \wedge p^y(1-p^y)\right) \cdot \left(\frac{1}{p^x + p^y - p^x p^y} + \frac{1}{1 - p^x p^y}\right).$$

□

1.3.3 Järjestikune joondus

Edasi uurime järjestikust joondust, mis on kujunenud Hammingu meetodi laiendusena. Kui vaatleme järjestikust joondust jada X põhjal, siis seisneb idee selles, et

1. joondame jada X esimese liikme jada Y esimese sama värvi liikmega,
2. joondame jada X teise liikme jada Y järgmise ettetuleva sama värvi liikmega nii, et eelnevaid joondamisi ei muudaks,
3. jätkame, kuni jadas Y ei leidu enam liikmeid, mida saaks joondada.

Seda meetodit saab rakendada ka blokkide korral, kuigi kahetähelise tähestiku tõttu ei erine see juba defineeritud bloki kaupa Hammingu meetodist.

Näide. Vaatleme jadasid

$$x = 0001100011110110000 \text{ ja}$$
$$y = 00001100001000110.$$

Nende jadade tähtede kaupa järjestikune joondamine (jada x põhjal) on *indel*'ite paigutuse täpsusega näiteks selline:

$$\begin{array}{l} 000\text{–}11000\text{–}1\text{–}\text{–}\text{–}11\text{–}10110000 \\ 00001100001000110\text{–}\text{–}\text{–}\text{–}\text{–}\text{–}\text{–}\text{–} \end{array}.$$

Defineerime järjestikuse meetodi skoori.

Definitsioon 18. *Olgu meil lõpmatud jadad X ja Y ning olgu $C_0 = 0$ ja*

$$C_k = \min\{i \geq 1 : Y_{S_{k-1}+i} = X_k\},\ k = 1, 2, \ldots,$$

kus $S_k = \sum_{i=1}^{k} C_i$. Jada $C_1, C_2, \ldots$ nimetame jada Y (tähe kaupa järjestikuse joonduse) sammude jadaks. Jada X põhjal tähtede kaupa järjestikuse joonduse skoor on

$$B(X)_n = \max\{k \geq 1 : S_k \leq n\},$$

Näide. Eelmise näite jadade tähtede kaupa järjestikuse joonduse skoor on $B(x)_{17} = 11$, sest $S_{11} = 17$, aga $S_{12} > 17$.

Järgnevalt tõestame tulemuse (jada X põhjal) järjestikuse joonduse keskmise skoori piirväärtuse kohta, mida tähistame γ_{CA} (ingl *consecutive alignment*).

Teoreem 8. *Olgu $X = X_1X_2\ldots$ ja $Y = Y_1Y_2\ldots$ lõpmatud, omavahel sõltumatud, Bernoulli jaotustega i.i.d. jadad, kusjuures $P(X_1 = 1) = p^x$ ja $P(Y_1 = 1) = p^y$. Siis on X ja Y sammude jada $C_1, C_2, \ldots$ i.i.d. jada ning järjestikuse joonduse keskmise skoori kohta kehtib p.k. koondumine*

$$\frac{B(X)_n}{n} \to \frac{1}{E(C_1)} = \frac{1}{\frac{p^x}{p^y} + \frac{1-p^x}{1-p^y}} =: \gamma_{CA},$$

Tõestus. Näitame, et jada $C_1, C_2, \ldots$ on *i.i.d.* jada. Selleks piisab näidata, et iga naturaalarvu $n, k_1, \ldots, k_n$ korral kehtib

$$P(C_1 = k_1, C_2 = k_2, \ldots, C_n = k_n) = P(C_1 = k_1)P(C_1 = k_2)\ldots P(C_1 = k_n).$$

Kuna iga naturaalarvu i ja j korral kehtib $P(C_i = j) =$

$$= \sum_{a=0,1} P(X_i = a, Y_{S_{i-1}+1} \neq a, Y_{S_{i-1}+2} \neq a, \ldots, Y_{S_{i-1}+(j-1)} \neq a, Y_{S_{i-1}+j} = a),$$

siis saame suvaliselt fikseeritud $n, k_1, \ldots, k_n \in \mathbb{N}$ korral, et

$$P(C_1 = k_1, C_2 = k_2, \ldots, C_n = k_n) =$$

$$= \sum_{a_1=0,1} \sum_{a_2=0,1} \ldots \sum_{a_n=0,1} P(X_1 = a_1, Y_1 \neq a_1, Y_2 \neq a_1, \ldots, Y_{k_1-1} \neq a_1, Y_{k_1} = a_1;$$

$$X_2 = a_2, Y_{k_1+1} \neq a_2, Y_{k_1+2} \neq a_2, \ldots, Y_{k_1+k_2-1} \neq a_2, Y_{k_1+k_2} = a_2; \ldots$$

$$X_n = a_n, Y_{k_1+\ldots+k_{n-1}+1} \neq a_n, \ldots, Y_{k_1+\ldots+k_n-1} \neq a_n, Y_{k_1+\ldots+k_n} = a_n) = \ldots$$

Kuna X ja Y on omavahel sõltumatud ning *i.i.d.* jadad ja eelmise tõenäosuse sees esineb iga juhuslik suurus ülimalt ühe korra, siis saame edasi kirjutada, kasutades kõigepealt sõltumatuse ja seejärel sama jaotuse omadusi, et

$$\ldots = \sum_{a_1=0,1} \sum_{a_2=0,1} \ldots \sum_{a_n=0,1} P(X_1 = a_1, Y_1 \neq a_1, Y_2 \neq a_1, \ldots, Y_{k_1-1} \neq a_1, Y_{k_1} = a_1)$$

$$P(X_2 = a_2, Y_{k_1+1} \neq a_2, Y_{k_1+2} \neq a_2, \ldots, Y_{k_1+k_2-1} \neq a_2, Y_{k_1+k_2} = a_2) \ldots$$

$$P(X_n = a_n, Y_{k_1+\ldots+k_{n-1}+1} \neq a_n, \ldots, Y_{k_1+\ldots+k_n-1} \neq a_n, Y_{k_1+\ldots+k_n} = a_n) =$$

$$= \sum_{a_1=0,1} \sum_{a_2=0,1} \ldots \sum_{a_n=0,1} P(X_1 = a_1, Y_1 \neq a_1, Y_2 \neq a_1, \ldots, Y_{k_1-1} \neq a_1, Y_{k_1} = a_1)$$

$$P(X_1 = a_2, Y_1 \neq a_2, Y_2 \neq a_2, \ldots, Y_{k_2-1} \neq a_2, Y_{k_2} = a_2) \ldots$$

$$P(X_1 = a_n, Y_1 \neq a_n, Y_2 \neq a_n, \ldots, Y_{k_n-1} \neq a, Y_{k_n} = a_n) =$$

$$P(C_1 = k_1)P(C_1 = k_2) \ldots P(C_1 = k_n).$$

Leiame definitsiooni põhjal juhusliku suuruse C_1 keskväärtuse:

$$E(C_1) = \sum_{i=1}^{\infty} i \cdot P(C_1 = i) =$$

$$= \sum_{i=1}^{\infty} i \cdot \Big(P(C_1 = i | X_1 = 1)P(X_1 = 1) + P(C_1 = i | X_1 = 0)P(X_1 = 0) \Big) =$$

$$= \sum_{i=1}^{\infty} i \cdot \Big(p^x (1-p^y)^{i-1} p^y + (1-p^x)(p^y)^{i-1}(1-p^y) \Big) =$$

$$= p^x \sum_{i=1}^{\infty} i \cdot \Big((1-p^y)^{i-1} p^y \Big) + (1-p^x) \sum_{i=1}^{\infty} i \cdot (p^y)^{i-1}(1-p^y) \Big) =$$

$$= p^x E(L_1^y | Y_1 = 0) + (1-p^x) E(L_1^y | Y_1 = 1).$$

Kuna nendel tingimustel on juhuslik suurus L_1^y geomeetrilise jaotusega, siis saame kokkuvõttes, et

$$E(C_1) = \frac{p^x}{p^y} + \frac{(1-p^x)}{1-p^y}.$$

Teoreemi 5 eeldused on jada $C_1, C_2, \ldots$ korral täidetud, seega teame, et

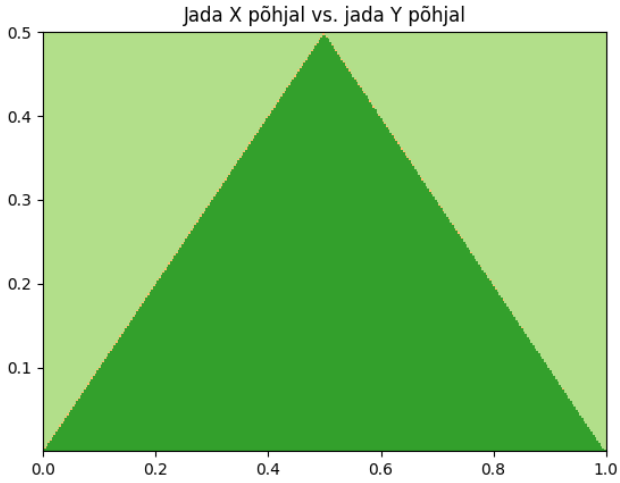
$$\frac{B(X)_n}{n} = \frac{M(n)}{n} \to \frac{1}{E(C_1)} = \frac{1}{\frac{p^x}{p^y} + \frac{1-p^x}{1-p^y}} =: \gamma^x_{CA}, \ p.k..$$

□

Kui joondamine käib jada Y järgi, siis tuleb tavaliselt erinev keskmise skoori piirväärtus, sest üldjuhul

$$\gamma^x_{CA} = \frac{1}{\frac{p^x}{p^y} + \frac{1-p^x}{1-p^y}} \neq \frac{1}{\frac{p^y}{p^x} + \frac{1-p^y}{1-p^x}} = \gamma^y_{CA}.$$

See võrdus kehtib vaid siis, kui $p^x = p^y$ või $1 - p^x = p^y$.



Joonis 1: Järjestikune joondus jada X põhjal vs. jada Y põhjal

Joonisel 1 on heleroheliseg märgistatud alad, kus $\gamma^x_{CA} > \gamma^y_{CA}$ ning tumeroheliseg alad, kus kehtib vastupidine võrratus.

2 Jadade võrdlemine

Selles peatükis uurime vaadeldud kolme meetodi käitumist parameetrite p^x ja p^y muutumisel. Alustuseks teeme mõnel erijuhul ennustusi selle kohta, kui hästi need meetodid töötavad, seejärel vaatame, kuidas nad peaksid teoreetiliselt käituma ning lõpetuseks teeme simulatsiooni, kus võrdleme meetodite skoore pikima ühisjadaga konkreetsete genereeritud jadade korral.

2.1 Meetodite käitumine erijuhtudel

Meie eesmärk on välja uurida, milline nendest meetoditest töötab antud parameetrite $p^x = P(X_1 = 1)$ ja $p^y = P(Y_1 = 1)$ korral kõige paremini ja milline neist lähendab kõige paremini optimaalset skoori kõigi väärtuste ulatuses. Võrdleme Bernoulli jaotusega, omavahel sõltumatuid *i.i.d.* jadasid X ja Y, kusjuures parameetrid p^x ja p^y muutuvad vahemikus $(0; 1)$, sest kui $p^x \in \{0, 1\}$ või $p^y \in \{0, 1\}$, siis on vähemalt üks jadadest determineeritud ja tulemus on triviaalne.

Parameetrite vahemikke saab veel kitsendada, sest jadade X ja Y kohti saame vabalt vahetada ning tähtede 0 ja 1 kohti samuti. Piisab vaadata vahemikke $p^x \in (0; 1)$ ja $p^y \in (0; 0{,}5)$, sest kui tahame teada, kuidas käituvad meetodid vahemikes $p^x \in (0; 0{,}5)$, $p^y \in (0{,}5; 1)$, piisab vahetada jadade X ja Y kohad ning vaadata vahemikke $p^x \in (0{,}5; 1)$, $p^y \in (0; 0{,}5)$. Samuti on ülejäänud kaks veerandit ruudus $p^x \in (0; 1)$, $p^y \in (0; 1)$ sümmeetrilised punkti $(0{,}5; 0{,}5)$ suhtes, kui defineerime parameetrid ümber kui $p^x = P(X_1 = 0), p^y = P(Y_1 = 0)$.

Vaatame nelja erijuhtu:

1. $p^x \approx 1, p^y \approx 0$ ehk jadas X on palju ühtesid, jadas Y on palju nulle.
2. $p^x \approx 0, p^y \approx 0$ ehk mõlemas jadas on palju nulle.
3. $p^x \approx 0{,}5, p^y \approx 0{,}5$ ehk mõlemas jadas on võrdselt ühtesid ja nulle.
4. $p^x \approx 1, p^y \approx 0{,}5$ ehk jadas X on palju ühtesid, aga jadas Y on võrdselt ühtesid ja nulle.

Analüüsime iga erijuhu korral kõigi kolme meetodi käitumist, arvutame kõigi parameetrite korral välja, milline meetod annab kõige parema tulemuse ning milline annab paremuselt teise. Kuna pikima ühisjada leidmiseks ei ole üldisel juhul valemit, siis saame teoreetiliselt võrrelda vaid meetodite omavahelisi suhteid.

1) Esimesel erijuhul koosneb jada X pikkadest 1-blokkidest ning lühikestest 0-blokkidest ja jada Y vastupidi lühikestest 1-blokkidest ning pikkadest 0-blokkidest. Sellisel juhul on parim, mida saab loota, see, et mõlema jada kõik lühemad blokid joondatakse vastavate pikkade blokkidega teisest jadast. Vaatleme jadasid

$$\begin{aligned} \text{x} &= 11111111111110111111110111111 \\ \text{y} &= 00000000010000000001000000000 \end{aligned}$$

Kui me joondame kõik jada x nullid mingite nullidega jadast y ja kõik jada y ühed mingite ühtedega jadast x, siis ei ole võimalik rohkem *match*'e saada ehk oleme saanud pikima ühisjada. Üks viis, kuidas seda saavutada on bloki kaupa Hammingu joondus, mis on praeguse näite puhul

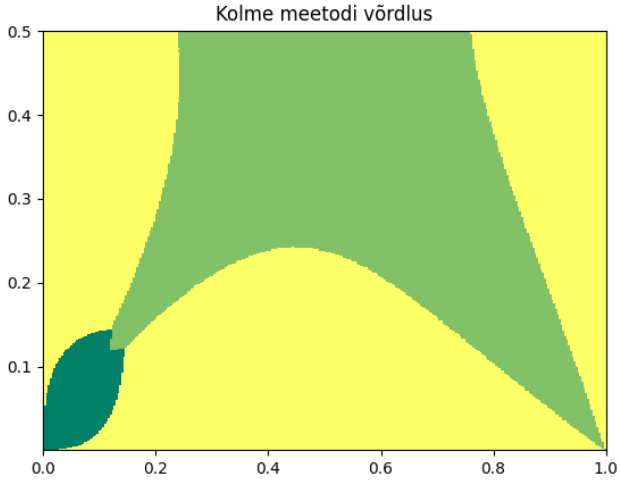
```
---------11111111111110---------1111111110--------111111
0000000001-----------00000000001--------000000000------ ’
```

mistõttu võib arvata, et bloki kaupa Hammingu joondus annab pikimale ühisjadale lähedase skoori. Järjestikune joondus jada x põhjal oleks aga

```
---------1----------1---------11111111110111111110111111
000000000100000000001000000000-------------------------
```

ehk joondasime ainult jada x ühed, aga mitte ühtegi jada y nulli. Seetõttu võiks arvata, et järjestikune joondus annab esimesel erijuhul halvema tulemuse kui bloki kaupa Hammingu joondus. Nagu teooriaosas mainitud, oleneb järjestikuse joonduse skoor sellest, kas see on leitud jada x põhjal või jada y põhjal, aga praeguse näite puhul annavad nad võrdse skoori.

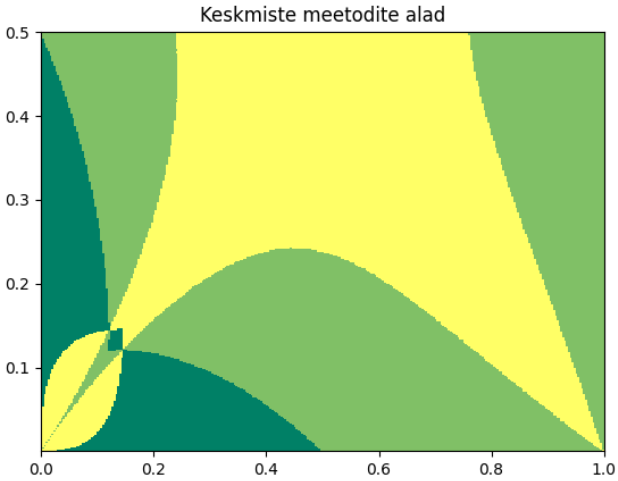
Tavaline Hammingu joondus annab hea skoori siis, kui võrreldavates jadades on ilma *indel*’eid lisamata palju *match*’e ja vähe *mismatch*’e, kuid ebasümmeetrilisel juhul on vastupidi palju *mismatch*’e ja vähe *match*’e. Meie näite puhul on *match*’e 4 ja *mismatch*’e 26, mistõttu Hammingu joonduse skoor on -22. Ka üldjuhul tuleb see skoor arvatavasti negatiivne.



Joonis 2: Iga meetodi alad, kus selle keskmise skoori piirväärtus on suurim

Joonisel 2 on iga meetodi jaoks märgistatud ala, kus selle keskmise skoori piirväärtus on teistest suurem. Tumerohelisega on tähistatud see ala tavalise Hammingu meetodi jaoks, heleroheliseg bloki kaupa Hammingu meetodi jaoks ning kollasega alad järjestikuse meetodi jaoks. Kui võrrelda jooniseid 1 ja 2, siis näeme, et vasak- ja parempoolne kollane ala on leitud järjestikuse joondusega jada X põhjal ning alumine kollane ala jada Y põhjal.

Intuitiivse analüüsi põhjal võis arvata, et bloki kaupa Hammingu joondus annab erijuhul (1) paremaid tulemusi kui järjestikune joondus, aga jooniselt 2 näeme, et see kehtib ainult kitsal alal. Lisaks sellele ütleb joonis 4, et sellel kitsal alal on bloki kaupa Hammingu joondus vaid mõnevõrra parem kui järjestikune joondus.



Joonis 3: Iga meetodi alad, kus selle keskmise skoori piirväärtus on teisel kohal

Joonisel 3 on iga meetodi jaoks kujutatud need alad, kus see meetod andis keskmise tulemuse ehk oli skoori piirväärtuse poolest arvuliselt teise kahe meetodi vahel. Kasutatud on samu värve, mis joonisel 2 ehk Hammingu meetod on tumeroheline, bloki kaupa Hammingu meetod on heleroheline ning järjestikune meetod on kollane.

2) Teine erijuht on võrreldavate jadade vahel sümmeetriline, aga tähtede suhtes ebasümmeetriline ning praegusel juhul on mõlemas jadas palju nulle ning vähe ühtesid. See tähendab, et ilma *indel*'eid lisamata on jadadel palju *match*'e ning vähe *mismatch*'e ehk *indel*'eid lisades on suurem võimalus, et kaotame *match*'i kui kaotame *mismatch*'i. Vaatleme näiteks jadasid

$$\begin{aligned} \text{x} &= 00000100000000010000000010000 \\ \text{y} &= 00000000010000000001000000000 \end{aligned}$$

Tavaline Hammingu joonduse skoor tuleb nende jadade jaoks $a_{match} - a_{mismatch} = 25 - 5 = 20$. Kuna mõlemas jadas on pikad 0-blokid, siis annab bloki kaupa Hammingu joondus

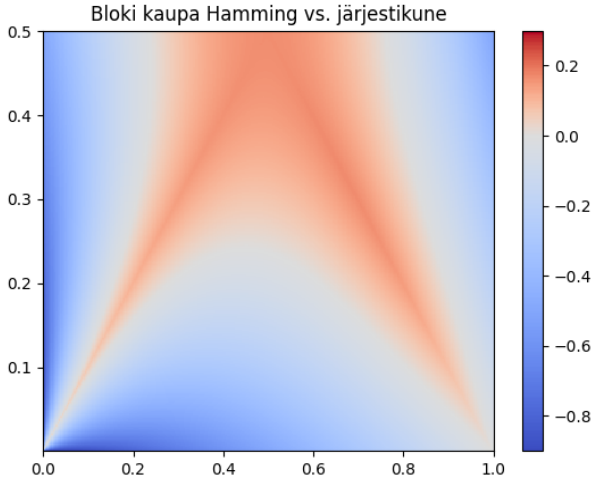
$$\begin{aligned} &00000\text{----}1000000000\text{-}100000000010000 \\ &000000000100000000001000000000\text{-----} \end{aligned}$$

siin suhteliselt hea skoori. Jadade pikkus on 30 ja *indel*'eid on kummalgi real 5 ehk skooriks saame 25. Jada x põhjal leitud järjestikune joondus on selle näite puhul täpselt sama, mis bloki kaupa Hammingu joondus, aga jada y põhjal on selleks

```
000001000000000100000000010000----------
00000-0000-----1000000000-0---1000000000
```

ning skoor on $30 - 10 = 20$. Seega võiks arvata, et ka üldjuhul annavad bloki kaupa Hammingu ja järjestikune joondus ligikaudu sama tulemuse ning tavaline Hammingu joondus on *mismatch*'ide tõttu nendest kehvem.

Jooniselt 2 näeme, et meie intuitiivne eeldus ei pidanud paika, sest all vasakus nurgas ei ole ala, kus bloki kaupa Hammingu joondus annaks parima tulemuse. Lisaks sellele näeme jooniselt 3, et erijuhul (2) on väga kitsas ala, kus bloki kaupa Hammingu joondus on skoori piirväärtuse poolest teisel kohal ja joonis 4 ütleb, et väljaspool seda kitsast ala on järjestikune joondus märgatavalt parem kui bloki kaupa Hammingu joondus. Tavalise Hammingu joonduse skoori piirväärtus on paljude väärtuste puhul esimesel kohal ehk parem, kui ennustasime.



Joonis 4: Vahe $\gamma_{BH} - (\gamma^x_{CA} \vee \gamma^y_{CA})$ väärtused

Joonisel 4 on kujutatud bloki kaupa Hammingu joonduse ja järjestikuse joonduse omavahelist suhet. Täpsemalt on punasega tähistatud alad, kus γ_{BH} on suurem kui $\gamma^x_{CA} \vee \gamma^y_{CA}$ ning sinisega alad, kus γ_{BH} on väiksem.

3) Nii jadade kui ka tähtede suhtes sümmeetrilisel juhul vaatame näiteks jadasid

```
x =00000001111100001111110000000
y =00000111111000111111000001111
```

Hammingu joonduse skooriks saame praegusel juhul $19 - 11 = 8$. Bloki kaupa Hammingu joondus on nende jadade puhul

```
000000011111–000011111100000000––––
00000––111111000–111111000000––1111
```

ning skoor on jadade pikkus miinus *indel*'ite arv ehk $30 - 5 = 25$. Järjestikune joondus jada x põhjal on

```
00000––––––00–11111–0000––11111100000000
0000011111100011111100000001111––––––––––
```

ning jada y põhjal

```
000000011111000011111100000000––––––––––––––––
00000––11111––––1–––––000–––––1111110000001111.
```

Nende skoorid on indelite arvu järgi vastavalt $30 - 10 = 20$ ja $30 - 16 = 14$. Selle näite abil võiks arvata, et sümmetrilisel juhul saab bloki kaupa Hammingu joondus üsna hea skoori, kuid järjestikune joondus võib terveid blokke vahele jätta, mistõttu tema skoor on halvem.

Seekord pidasid ennustused paika, sest jooniste 2 ja 3 põhjal on erijuhul (3) bloki kaupa Hammingu joondus kõige edukam ning järjestikune joondus on teisel kohal.

4) Viimane erijuht on esimesega sarnane, aga praegu jõuame pikima ühisjada pikkusele kõige lähemale siis, kui saame jada X kõik lühikesed 0-blokid joondatud keskmise pikkusega 0-blokkidega jadast Y ning jada Y kõik keskmise pikkusega 1-blokid joondatud pikkade 1-blokkidega jadast X. Kuna jadal Y on 1-blokid pikemad kui need olid esimesel juhul, siis peaks pikim ühisjada ise olema ka pikem. Vaatleme järgimisi jadasid

```
x =11111111111110111111110111111
y =11111000011111000001111000111.
```

Hammingu joonduse skooriks saame $a_{match} - a_{mismatch} = 16 - 14 = 2$. Bloki kaupa Hammingu joondus tuleb seekord

```
11111111111110–––1111111110––––111111––––––
11111––––––––000011111––––0000011111–000111
```

ning skooriks saame *indel*'ite järgi $30 - 11 = 19$. Näeme, et jada y järgi leitud järjestikune joondus annab skooriks kõigest 7 ning leiame järjestikuse joonduse jada x põhjal, milleks on

```
11111----11111-----111--0--1111111110111111
1111100001111100000111110001111-----------'
```

skooriga 17.

Hammingu joondus annab seekord tõenäoliselt parema skoori kui erijuhul (3), sest jadas X on rohkem ühtesid, mistõttu *mismatch*'e on vähem. Teised kaks meetodit andsid praegu umbes sama skoori, aga võib arvata, et bloki kaupa Hammingu meetod annab pikemate jadade korral halvema skoori. Jadas X on pikemad 1-blokid kui jadas Y, mistõttu on neid vähem kui jadas Y. See aga tähendab, et kui me joondame blokkide haaval, siis saavad jadast Y blokid enne otsa kui jadast X ehk mingi osa 1-blokke jääb jadast Y joondamata. Jada X põhjal leitud järjestikuse joonduse puhul saame aga peaaegu kõik 1-blokid jadast Y joondatud, mistõttu peaks see andma parema skoori.

Näeme jooniste 2 ja 4 põhjal, et järjestikune joondus annabki erijuhul (4) parema tulemuse kui bloki kaupa Hammingu meetod.

2.2 Simulatsioonid

2.2.1 Simulatsiooni kirjeldus

Tahame uurida, kui häid tulemusi annavad vaadeldavad meetodid erinevate parameetrite väärtuste korral, kuid teooriaosas leitud keskmiste skooride piirväärtused γ_H, γ_{BH} ning $\gamma^x_{CA}, \gamma^y_{CA}$ ei anna meile infot selle kohta, mis on antud parameetrite korral ülemine tõke võrreldavate jadade skoorile ehk mis on optimaalne skoor. Teooriaosas näitasime, et kui kasutada sarnasusfunktsiooni

$$s(x_i^*, y_i^*) = \begin{cases} 1, & \text{kui } x_i^* = y_i^* \ (match), \\ -1, & \text{kui } - \neq x_i^* \neq y_i^* \neq - \ (mismatch), \\ 0, & \text{kui } x_i^* = - \text{ või } y_i^* = -, \end{cases}$$

siis tähendab optimaalse skoori leidmine ühtlasi pikima ühisjada pikkuse leidmist. Viimase jaoks on vaja teada ühte pikimat ühisjada, mille saab leida laialdaselt kasutatava Needleman-Wunschi algoritmi abil (Barder *et al.*, 2012). Simulatsiooni idee seisneb järgnevas:

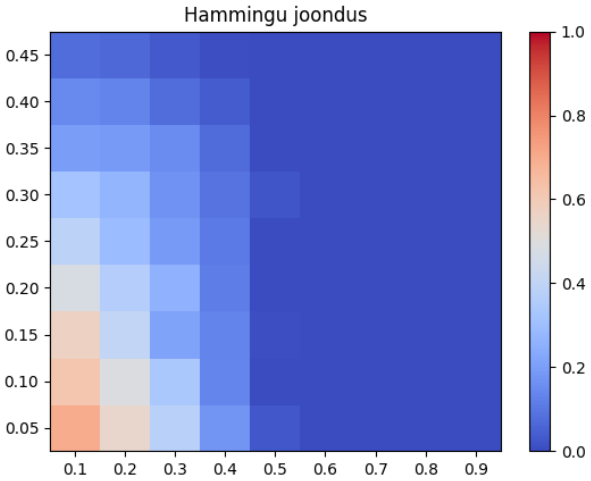
- Koostame parameetrite p^x ja p^y jaoks võrgustiku, mis katab terve vaadeldava ala, nimelt $p^x \in \{0{,}1; 0{,}2; \ldots ; 0{,}9\}$ ja $p^y \in \{0{,}05; 0{,}1, \ldots ; 0{,}45\}$.
- Iga punkti (p_1^x, p_1^y) korral genereerime juhuslikult jadad x ja y pikkusega 2500. Jada x elemendid on saadud Bernoulli katsega, kus tähe 1 saamise tõenäosus on p_1^x ning jada y elementidel on tähe 1 saamise tõenäosus p_1^y.

- Iga genereeritud jadade paari jaoks leiame nende Hammingu joonduse skoori, bloki kaupa Hammingu joonduse skoori, järjestikuse joonduse skoori jada x põhjal ja jada y põhjal ning nende pikima ühisjada pikkuse ehk optimaalse skoori Needleman-Wunschi algortmi abil. (Graafikutes kasutame kas jada x või jada y põhjal leitud järjestikuse joonduse skoori, olenevalt sellest, kumb tuli suurem.)
- Saadud info salvestame failidesse, et seda hiljem joonistel kujutada.

Oleksime valinud tihedama punktide võrgustiku ning genereerinud pikemaid jadasid, aga Needleman-Wunschi algoritmi jooksutamise mahukus seadis nendele ülemise piiri. Needleman-Wunschi algoritmi koodi saime Internetist ning kohandasime seda praegusel juhul skoori leidmiseks (Slowikowski, 2020). Simulatsiooni tööks oli tarvis programmeerida ka kolme meetodi skoori arvutamise funktsioonid, mis on koos simulatsiooniga toodud lisas 1.

2.2.2 Simulatsiooni analüüs

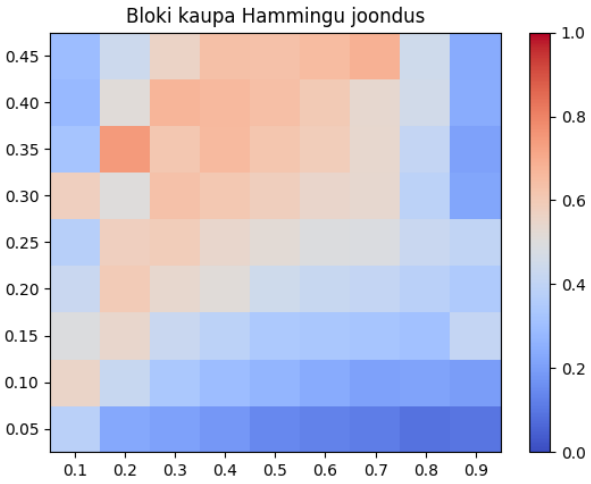
Järgneva nelja joonisega on välja toodud simulatsiooni tulemused, kus igal joonisel on toodud ühe meetodi skoori suhe jadade pikkusega $n = 2500$. Jooniselt 5 näeme, et Hammingu joondus töötab kõige paremini tõepoolest erijuhul (2) ning selle ümbruses, nagu jooniste 2 ja 3 pealt saab näha.



Joonis 5: Simulatsiooni tulemus Hammingu meetodi jaoks

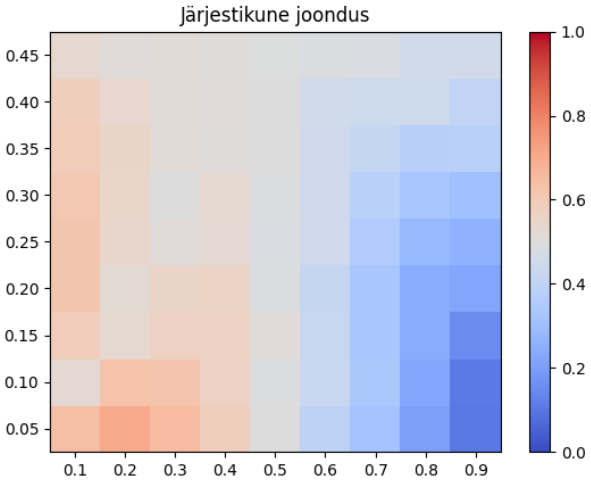
Nagu ennustasime, on skoor erijuhul (1) negatiivne, kuid jätsime Hammingu joonduse graafikul negatiivsed väärtused välja, et seda saaks hõlpsamalt teiste graafikutega võrrelda. Toome eraldi välja, et Hammingu joonduse tulemus punktis

$(0{,}9; 0{,}05)$ oli $-0{,}7312$ ning punktis $(0{,}9; 0{,}45)$ oli selleks $-0{,}0976$ ehk skoor erijuhul (4) oli parem kui erijuhul (1), nagu ennustasime.



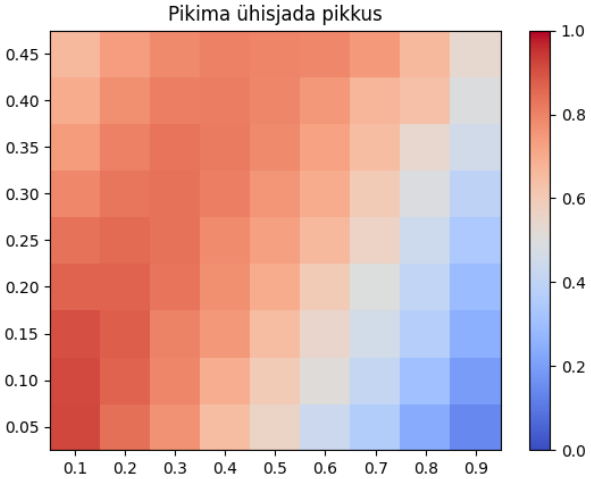
Joonis 6: Simulatsiooni tulemus bloki kaupa Hammingu meetodi jaoks

Simulatsioon kinnitab joonistega 6 ja 7, et bloki kaupa Hammingu joondus annab erijuhul (3) paremaid tulemusi kui järjestikune joondus ning et erijuhul (2) on samuti kitsas ala, kus bloki kaupa Hammingu joondus on parem. Lisaks näeme, et genereeritud jadade puhul jäi kehtima see, et punktide $(0{,}5; 0)$ ja $(0; 0{,}25)$ ümbruses annab järjestikune joondus paremaid tulemusi.



Joonis 7: Simulatsiooni tulemus järjestikuse meetodi jaoks

Jooniselt 8 näeme, et genereeritud jadade puhul käitub pikima ühisjada pikkus nii, nagu võiks arvata. Seal, kus jadad on sümmeetrilised ehk $p^x \approx p^y$, on optimaalne skoor suurem ja seal, kus jadad on ebasümmeetrilised ehk $1 - p^x \approx p^y$, on optimaalne skoor väiksem. Näeme ka seda, et erijuhul (1) on pikima ühisjada pikkus väiksem kui erijuhul (4), nagu ennustasime. Lisaks näeme joonist 8 eelmistega võrreldes, et järjestikune joondus imiteerib pikima ühisjada pikkuse graafikut kõige paremini ehk üldjuhul tasub nende kolme meetodi vahel valides eelistada järjestikust joondust (nii jada X kui ka jada Y põhjal, olenevalt parameetritest). Erijuhul (2) tasub valida Hammingu joondus, sest see annab kõige parema tulemuse ja ei nõua *indel*'ite lisamist ja erijuhul (3) tuleks eelistada bloki kaupa Hammingu joondust, sest see annab pikima ühisjada pikkusele lähedasema skoori kui teised. Seega igal meetodil on vähemalt mingi erijuht, kus see annab võrreldes teistega parema skoori.



Joonis 8: Simulatsiooni tulemus pikima ühisjada pikkuse jaoks

Kokkuvõte

Töö käigus vormistasime ranged tõestused iga meetodi keskmise skoori piirväärtuse koondumiseks ning kontrollisime nende kehtivust simulatsiooni abil. Ühtlasi saime simulatsiooni käigus natuke parema arusaama sellest, kuidas pikima ühisjada pikkus parameetritest (p^x, p^y) sõltub.

Töö tulemusena saame vastata ka püstitatud uurimisküsimustele. Leidsime, et kõik kolm suboptimaalset meetodit annavad parameetrite (p^x, p^y) mingite väärtuste korral paremaid tulemusi kui teised kaks meetodit ehk mitte ühtegi neist ei tasu kõrvale jätta. Lisaks saime simulatsiooni abil teada, et vaadelduna terves parameetrite väärtuste ulatuses, lähendab järjestikune joondus nendest meetoditest kõige paremini pikima ühisjada pikkust.

Meetodite väikse arvu tõttu ei anna käesolev töö head ülevaadet kõigist suboptimaalsetest meetoditest, mida on uuritud (Klement ja Lember, 2017; Barder *et al.*, 2012). Edasi saaks võrrelda nende meetodite tööd rafineeritumate suboptimaalsete meetoditega, mis annavad arvatavasti paremaid tulemusi. Kuna kitsendasime selle töö kahetähelisele tähestikule, siis saaks edaspidi uurida ka kõigi nende meetodite käitumist kõigis parameetrite väärtustes ja suurema tähestiku korral.

Kasutatud allikad

Barder, S., J. Lember, H. Matzinger ja M. Toots (2012). „On Suboptimal LCS-Alignments for Independent Bernoulli Sequences with Asymmetric Distributions“. *Methodol Comput Appl Probab* 14, lk. 357–382.

Klement, R. ja J. Lember (2017). „On expected score of cellwise alignments“. *ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA* 21.1, lk. 141–165.

Slowikowski, K. (2020). *A simple version of the Needleman-Wunsch algorithm in Python*. URL: https://gist.github.com/slowkow/06c6dba918\0d013dfd82bec217d22eb5 (vaadatud 09.05.2022).

Toots, M. (2008). „Chvatal-Sankovi konstandi hindamine simulatsioonide abil“. Bakalaureusetöö. Tartu Ülikool.

Toots, M. (2012). „Suboptimal Alignments and Similarity of Random Sequences“. Magistritöö. Tartu Ülikool.

Lisa 1. Skooride arvutamise programmid ja simulatsioon

```
import numpy as np
import pandas as pd

def gener(px, py, n): # jadade x ja y juhuslik genereerimine
    x = np.random.binomial(1, px, n)
    y = np.random.binomial(1, py, n)
    return x, y


def ham_skoor(x, y, n): # Hammingu joonduse skoori leidmine
    vahe = abs(x - y)
    mismatch = sum(vahe)
    match = n - mismatch
    return match - mismatch


def blokham_skoor(x, y, indeks):
    # jätame vajadusel y-st esimese vale bloki ära
    if indeks == 2: # rekursiivsuse tõkestamiseks kahele korrale
        return 0
    x_esimene_täht = x[0]
    y_algus = 0
    # leiame jadast y esimese tähe, mis on x esimese tähega sama
    for i in range(0, len(y), 1):
        if y[i] == x_esimene_täht:
            y_algus = i
            break         # kui terve jada y on üks blokk, siis
        if i == len(y)-1: # vahetame x ja y ära ning proovime uuesti
            indeks += 1 # indeks ütleb meile, kui mõlemad jadad
            return blokham_skoor(y, x, indeks) # on ainult üks blokk
    y = y[y_algus:] # saime vajaliku y-i
    skoor = 0
    x_blokk = 0 # salvestame x-i ja y-i käesoleva bloki pikkused
    y_blokk = 0
    x_eelmine = 0 # salvestame x-i ja y-i eelmise bloki lõpu indeksi
    y_eelmine = 0
    while (x_eelmine < len(x)) & (y_eelmine < len(y)):
        for j in range(x_eelmine, min(len(x)+1, len(y)+1), 1):
```

```
            try:
                if x[j] != x[j+1]: # otsime bloki lõppu
                    x_blokk = (j+1) - x_eelmine
                    x_eelmine = j+1
                    break
            except IndexError: # kui jõudsime jada x lõppu
                x_eelmine = len(x)+1 # et while-tsükkel lõppeks
                x_blokk = 0 # et ei liidetaks rohkem skoori juurde
                break
        for j in range(y_eelmine, min(len(x), len(y)), 1):
            try:
                if y[j] != y[j+1]: # otsime bloki lõppu
                    y_blokk = (j+1) - y_eelmine
                    y_eelmine = j+1
                    break
            except IndexError: # kui jõudsime jada y lõppu
                y_eelmine = len(y)+1 # et while-tsükkel lõppeks
                y_blokk = 0 # et ei liideks rohkem skoori juurde
                break
        skoor += min(x_blokk, y_blokk)
    return skoor


def jarj_skoor(x, y): # järjestiku joonduse skoori leidmine
    skoor = 0
    y_eelmine = 0 # salvestame eelmise y indeksi
    näitaja = 0 # kasutame siis, kui jada y saab läbi vaadatud
    for i in range(0, len(x), 1):
        if näitaja == 1: # kui jõuame jada y lõppu, lõpetame
            break
        for j in range(y_eelmine, len(y)+1, 1):
            try:
                if y[j] == x[i]: # otsime jadast y järgmist
                    skoor += 1   # kattuvat tähte
                    y_eelmine = j+1
                    break
            except IndexError: # jõudsime jada y lõppu
                näitaja = 1 # et välimine tsükkel lõppeks
    return skoor

def nw(x, y, match=1, mismatch=1, gap=0): # kood on veebilehelt
    return # skoor
```

```
# simulatsioon

x0 = np.linspace(0.1, 0.9, 9) # koostame võrgustiku
y0 = np.linspace(0.05, 0.45, 9)
ham_maatriks = []   # tühjad järjendid, kuhu hakkame
bham_maatriks = []  # simulatsiooni tulemusi lisama
jarjx_maatriks = []
jarjy_maatriks = []
nw_maatriks = []
recur_index = 0
n = 2500 # genereeritavate jadade pikkus

for i in range(0, len(x0), 1):
    ham_maatriks.append([])
    bham_maatriks.append([])
    jarjx_maatriks.append([])
    jarjy_maatriks.append([])
    nw_maatriks.append([]
    # genereerime juhuslikult kaks jada ning
    # leiame skoorid ja lisame need maatriksisse
    for j in range(0, len(y0), 1):
        x, y = gener(x0[i], y0[j], n)
        ham_maatriks[i].append(ham_skoor(x, y, n)/n)
        bham_maatriks[i].append(blokham_skoor(x, y, recur_index)/n)
        jarjx_maatriks[i].append(jarj_skoor(x, y)/n)
        jarjy_maatriks[i].append(jarj_skoor(y, x)/n)
        nw_maatriks[i].append(nw(x, y)/n)

jarjmax_maatriks = np.maximum(jarjx_maatriks, jarjy_maatriks)
# kasutame jada x põhjal ning jada y põhjal leitud
# skoori maksimumi

df = pd.DataFrame(data=ham_maatriks) # salvestame info failidesse
df.to_csv('ham.csv', header=False, index=False)
df = pd.DataFrame(data=bham_maatriks)
df.to_csv('bham.csv', header=False, index=False)
df = pd.DataFrame(data=jarjmax_maatriks)
df.to_csv('jarjmax.csv', header=False, index=False)
df = pd.DataFrame(data=nw_maatriks)
df.to_csv('nw.csv', header=False, index=False)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristjan Vimm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Suboptimaalsed meetodid jadade sarnasuse võrdlemiseks kahetähelise tähestiku puhul, mille juhendaja on Jüri Lember, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristjan Vimm
10.05.2022