Tampereen yliopisto

Jennifer Nguyen

MARKOVIN PIILOPROSESSI SEKVENSSIANALYYSISSÄ

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Helmikuu 2020

Tiivistelmä

Jennifer Nguyen: Markovin prosessi sekvenssianalyysissä
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastotieteen tutkinto-ohjelma
Helmikuu 2020

Tutkielma esittelee Markovin piiloprosessin ja sen merkityksen sekvenssianalyysissä. Sekvenssianalyysillä halutaan määrittää jonkin tietyn sekvenssin emäsjärjestys, ja Markovin piiloprosessilla voidaan määrittää todennäköisyysjakauma jokaiselle sekvenssille ja lopulta valita se järjestys, joka on kaikista todennäköisin toteutua. Markovin piiloprosessi pohjautuu Markovin prosessiin, joka käsitellään tutkielmassa yhdessä todennäköisyyslaskennan perusaskelien kanssa.

Avainsanat: sekvenssianalyysi, todennäköisyyslaskenta, Markovin prosessi ja Markovin piiloprosessi.

Sisältö

1 Johdanto

Markovin prosessi tai malli on tilastotieteen prosessi. Prosessin tuleva tila riippuu ainoastaan prosessin nykyisestä tilasta. Markovin prosessista voidaan johtaa Markovin piiloprosessi, jonka tarkoituksena on laskennallisessa biologiassa löytää ”piilossa” olevia sekvenssejä. Markovin piiloprosessia on esitelty tilastollisessa kirjallisuudessa jo vuodesta 1966 lähtien. 1970-luvulla Markovin piiloprosessia on käytetty puheentunnistamisessa, mikä on ollut aikaisimpia piiloprosessin toimintoja eimatemaattisessa konteksissa. Nykypäivänä Markovin piiloprosessia on käytetty paljon bioinformatiikassa ja laskennallisessa biologiassa. Monet algoritmit, joita käytetään geenien tunnistamiseen pohjautuvat Markovin prosesseihin ja sen laajennoksiin.

Markovin piiloprosessi on erityisesti laskennallisessa biologiassa käytetty todennäköisyysmalli. Usein sekvensoinnissa halutaan määrittää ja muodostaa sekvenssielementtien tarkka järjestys. Markovin piiloprosessi on looginen tapa tehdä lineaaristen sekvenssien todennäköisyysmalleista päätelmiä. Markovin piiloprosessi on keskeinen ratkaisu laskennallisessa sekvenssianalyysissä. Tutkielman päälähteenä on käytetty Vidyasagarin teosta Hidden Markov processes: theory and applications to biology [4].

Työ esittelee kolme aihetta. Ensimmäisessä luvussa esitellään aiheen biologista taustaa siten että, biologiaan perehtymätön lukija pystyy seuraamaan. Toisessa luvussa tarkastellaan syvemmin matemaattisia määritelmiä ja kaavoja, jotka perustuvat Markovin prosessin laskemiseen. Lopuksi esitellään Markovin prosessi ja Markovin piiloprosessi ja kuinka jälkimmäistä voidaan hyödyntää sekvenssianalyysissä.

2 Biologinen konteksti

Geenit eli perintötekijät ovat DNA-molekyylin toiminnallisia osia. Geenit ohjaavat proteiinin toimintaa ja tuotantoa. Proteiinit, nukleiinihapot, lipidit ja hiilihydraatit ovat elämälle olennaisia orgaanisia molekyylejä. Proteiinit eli peptidit ovat aminohappoketjuja, jotka syntyvät solujen rakenneosina ja entsyymeinä.

Proteiinisynteesissä DNA-ketjuun on koodattuna informaatiota, joka kopioidaan RNA-molekyyliin. Silmukointi on proteiinisynteesin introinijaksojen poistamisvaihe, jossa poistetaan RNA:n esiasteesta jaksot, jotka eivät koodaa geenin proteiinia. Tällöin jäljelle jää vain geenin koodavat osat eli eksonit. RNA-molekyylistä informaatio siirtyy rakennettavan aminohappojärjestyksen ohjeena ribosomiin, jossa aminohappoketjuja ja RNA-komponentteja yhdistetään peptidisidoksilla polypeptideiksi eli proteiineiksi. Ribosomi on pieni soluelin, joka koostuu RNA:sta ja proteiineista.

DNA-ketju on kierteinen tikapuurakenteinen molekyyli, joka koostuu kahdesta nukleiinihapporihmasta. RNA koostuu yhdestä nukleiinihapporihmasta. DNA-ketju sisältää neljää eri emästä: adeniinia (A), sytosiinia (C), guaniinia (G) ja tymiiniä(T). DNA-ketjun pituutta mitataan emäspareina. Nukleotidi eli DNA:n rakennusyksikkö koostuu emäs-, fosfaatti ja sokeriosasta. DNA-sekvenssiksi tai pelkäksi sekvenssiksi kutsutaan siis emäsjärjestystä.

DNA-sekvenvoinnilla yritetään määrittää nukleotidien tarkka järjestys DNA-molekyylissä. Määritetään neljän emäksen järjestys geneettisen koodin selvittämiseksi. DNA-sekvensointi mahdollistaa paitsi ihmisen myös monien muiden lajien koko DNA:n eli genomin kartoituksen [1, s. 47-57].

3 Todennäköisyyden esittely

Seuraavaksi esitellään matemaattisia määritelmiä, lauseita ja kaavoja, jotka ovat olennaisia laskennallisen biologian määrittämisen kannalta. Suurin osa laskennallisista menetelmistä perustuu tapahtumiin, jotka ovat riippumattomia toisistaan. Esimerkiksi tietyn nukleotidin esiintyminen tietyssä DNA-jakson paikassa on riippumaton jakson muiden nukleotidien paikasta. Vastaavasti toiset mallit sallivat riippuvuuden, jolloin tietyn nukleotidin esiintymisen todennäköisyys tietyssä paikassa riippuu muun sekvenssin asettamasta ehdosta. Näin ollen esitetään määritelmästä johdettavat kaavat Markovin prosessille. Tämän luvun määritelmät mukailevat Vidyasagarin teoksen määritelmiä [4].

3.1 Satunnaismuuttujat ja satunnaisvektorit

Määritelmä 3.1. Olkoon S_n on määritelty kaavalla

$$S_n = \left\{ \boldsymbol{v} \in \mathbb{R}^n \;\middle|\; \boldsymbol{v}_i \geq 0 \; \forall i, \sum_{i=1}^{n} \boldsymbol{v}_i = 1 \right\}$$

eli S_n on joukko ei-negatiivisia vektoreita, joiden komponenttien summa on yksi.

Määritelmä 3.2. Olkoon $\mathbb{A} = \{a_1, a_2, \ldots, a_n\}$ on äärellinen joukko. Nyt joukon $\mathbb{A}$ todennäköisyysjakauma on mikä tahansa vektori $\mu \in S_n$. Nyt satunnaismuuttujan X todennäköisyys $X = a_i$ on

$$P\{X = a_i\} = \mu_i.$$

Lause 3.3. *Olkoon $\mathbb{A}$ äärellinen joukko ja μ sen todenäköisyysjakauma. Merkitään joukon $\mathbb{A}$ todennäköisyysmittaa notaatiolla P. Silloin seuraavat ominaisuudet ovat voimassa.*

1. $0 \leq P(A) \leq 1 \quad \forall A \subseteq \mathbb{A}$.
2. $P(\emptyset) = 0$ *ja* $P(A) = 1$.
3. *Jos A ja B ovat erillisiä joukkoja, niin*

$$P(A \cup B) = P(A) + P(B).$$

4. *Jos $A \subseteq B$, niin $P(A) \leq P(B)$.*

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

[2, s. 14].

Määritelmä 3.4. Olkoot $A = \{a_1, a_2, \ldots, a_n\}$ ja $B = \{b_1, b_2, \ldots, b_n\}$ satunnaismuuttujien X ja Y arvojoukkoja. Joukkojen A ja B karteesinen tulo $A \times B$ on

$$A \times B = \{\, (a_i, b_j) \mid a_i \in A,\, b_i \in B \,\}.$$

Lisäksi yhdistetyn satunnaismuuttujan $Z = (X, Y)$ arvojoukko on $A \times B$. Määritellään

$$P\{Z = (a_i, b_j)\} = P\{X = a_i \wedge Y = b_j\} \quad \forall a_i \in X \text{ ja } b_j \in Y.$$

3.2 Satunnaismuuttujien funktiot ja odotusarvo

Olkoon joukon $A = \{a_1, a_2, \ldots, a_n\}$, todennäköisyysmitta P, missä μ on todennäköisyysjakauma. Olkoon f kuvaus joukolta A joukkoon B. Koska A on äärellinen, niin myös $\{f(a_1), f(a_2), \ldots, f(a_n)\}$ on äärellinen. Nyt $P\{X = x_i\} = \mu_i$ ja merkitään $B = \{b_1, b_2, \ldots, b_n\}$ kuvaamaan kuvauksen $f(x)$ kaikkia tuloksia. Tällöin saadaan

$$P\{f(x) = b_j\} = \sum_{a_i \in f^{-1}(b_j)} \mu_i.$$

Toisin sanoin todennäköisyys, että $f(x) = b_j$ on kaikkien b_j:n käänteiskuvien todennäköisyyksien summa.

Esimerkki 3.5. Olkoon $R = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ joukko DNA nukleotideja ja $R = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ joukko RNA nukleotideja. Transkriptiossa DNA:n tymiini (T) korvataan RNA urasiililla (U). Translaatiossa RNA:n nukeiinikolmikko eli kodoni muuttuu aminohapoksi tai lopetuskodoniksi. Geneettiseksi koodiksi kutsutaan kuvausta, missä jokainen kodoni muutetaan aminohapoksi. Merkitään symbolilla $A = \mathcal{R}^3$ kodonien joukkoa, jonka määrä on $4^3 = 64$, ja olkoon B aminohappojen ja STOP-kodonin joukko, joiden lukumäärä on 21. Geneettinen koodi voidaan ajatella funktioksi $f\colon A \to B$.

Määritelmä 3.6. Olkoon $A = \{a_1, a_2, \ldots, a_n\} \subseteq \mathbb{R}$ äärellinen joukko. Merkitään μ kuvaavaan satunnaismuuttujan todennäköisyysjakaumaa. Nyt satunnaismuuttujan X odotusarvo on

$$\mathrm{E}\, X = \sum_{i=1}^{n} a_i \mu_i = \sum_{i=1}^{n} a_i\, P\{X = a_i\}.$$

Määritelmä 3.7. Funktioiden odotusarvoa merkitään notaatiolla $\mathrm{E}[f, P]$ ja määritellään

$$\mathrm{E}[f, P] := \sum_{i=1}^{n} f(a_i)\mu_i = \sum_{i=1}^{n} f(a_i)\, P\{X = a_i\}.$$

Satunnaismuuttujan varianssi on $\mathrm{E}[(X - (E))^2, P]$ ja keskihajonnaksi kutsutaan neliöjuurta varianssista.

3.3 Riippumattomuus ja ehdollinen jakauma

Määritelmä 3.8. Olkoot X ja Y satunnaismuuttujia, joiden arvojoukot ovat A ja B. Satunnaismuuttujien X ja Y sanotaan olevan *riippumattomia*, jos

$$P\{X = a_i \wedge Y = b_j\} = P\{X = a_i\} \cdot P\{Y = b_j\}.$$

Määritelmä 3.9. Olkoon X ja Y satunnaismuuttujia joiden äärelliset arvojoukot ovat A ja B. Satunnaismuuttujan X *ehdollinen todennäköisyys* satunnaismuuttujan Y suhteen määritellään

$$P\{X = a_i \mid Y = b_j\} = \frac{P\{X = a_i \wedge Y = b_j\}}{P\{Y = b_j\}}.$$

Määritelmä 3.10. Olkoot X, Y ja Z satunnaismuuttujia, joiden arvojoukot ovat $A = \{a_1, a_2, \ldots, a_n\}$, $B = \{b_1, b_2, \ldots, b_n\}$ ja $C = \{b_1, b_2, \ldots, b_n\}$. Nyt X ja Y ovat ovat riippumattomia ehdolla Z, jos $\forall a_i \in A, b_j \in B \, ja \, c_k \in C$,

$$P\{X = a_i \wedge Y = b_j \mid Z = c_k\} = P\{X = a_i \mid Z = c_k\} \cdot P\{Y = b_j \mid Z = c_k\}.$$

Lause 3.11. *Bayesin kaava. Olkoot X ja Y satunnaismuuttujia, joiden arvojoukot ovat A ja B. Näin ollen*

$$P\{X = a_{ij} \mid Y = b_j\} = \frac{P\{Y = b_j \mid X = a_j\} \cdot P\{X = a_i\}}{P\{Y = b_j\}}.$$

Todistus. Määritelmän 3.9 mukaan satunnaismuuttujan X todennäköiyys ehdolla Y on

$$P\{X = a_i \mid Y = b_j\} = \frac{P\{X = a_i \wedge Y = b_j\}}{P\{Y = b_j\}}.$$

Toisaalta määritelman 3.9 mukaan

$$P\{X = a_i \wedge Y = b_j\} = P\{X = a_i\}\, P\{Y = b_j \mid X = a_i\}.$$

Kun sijoitetaan ehdolliseen todennäköisyyteen tarvittavat tiedot, niin saadaan

$$P\{X = a_{ij} \mid Y = b_j\} = \frac{P\{X = a_j \wedge Y = b_j\}}{P\{Y = b_j\}} = \frac{P\{Y = b_j \mid X = a_i\} \cdot P\{X = a_i\}}{P\{Y = b_j\}}.$$

□

4 Markovin piiloprosessi

Tässä luvussa keskitytään aiheen pääasiaan eli Markovin piiloprosessiin. Markovin piiloprosessi (engl. ”hidden Markov proscess”, HMP) avulla voidaan tunnistaa geenejä tai muita mielenkiintoisia ominaisuuksia annetusta sekvenssistä yksinkertaista Markovin prosessia tehokkaammin. Osassa 4.1 tutkitaan Markovin prosessia, joka toimii pohjana Markovin piiloprosessille. Osassa 4.2 tarkastellaan Markovin piiloprosessin ominaisuuksia ja useita eri esitysmuotoja. Viimeisessä osassa 4.3 tutkitaan piiloprosessin roolia sekvenssianalyysissä. Tämä luku on tehty mukaillen Vidyasagarin [4] esitystapaa. Lisäksi viimeinen osa mukailee Nykterin [3] luennon esitystapaa.

4.1 Markovin ketju ja stokastiset mallit

Pohjaudutaan ensin Markovin prosessiin, josta johdetaan Markovin piiloprosessi. Markovin prosessissa tai mallissa tuleva tila riippuu ainoastaan prosessin nykytilasta. Voidaan esittää jokin tila *Markovin ketjuna* $(X_0, X_1, \ldots, X_t)$, missä jokainen X_0^t oletetaan kuuluvan äärelliseen joukkoon. Merkitään jatkossa lyhyesti $X_0^t = X_0, X_1, \ldots, X_t$.

Määritelmä 4.1. Olkoon $\mathbb{A}$ on äärellinen joukko. Stokastinen prosessi $\{X_t\}_{t=0}^{\infty}$ on Markovin prosessi, jos kaikilla t:n arvoilla $t \geq 1$ ja sekvensseillä $y_0, y_1, \ldots, y_{t-1}, y_t \in \mathbb{A}^{t+1}$ pätee, että

$$P\{X_t = y_t \mid X_0 = y_0, \ldots, X_{t-1} = y_{t-1}\} = P\{X_t = y_t \mid X_{t-1} = y_{t-1}\}.$$

Kaavaa kutsutaan *Markovin ominaisuudeksi*.

Esimerkki 4.2. Otetaan esimerkiksi nukleiinihappojono. Jokaisessa vaiheessa nukleiinihapon jakauma riippuu ainoastaan välittömästä edellisestä nukleiinihaposta, eikä sitä edeltävistä. Markovin prosessissa ehdolliset jakaumat $P(X_t \mid X_t - 1)$ ovat samat kaikilla, ajanhetkellä t.

Määritelmä 4.3. Stokastinen prosessi $\{X_t\}$ on Markovin prosessi, jos kaikilla $y_0^t \in \mathbb{A}$ pätee

$$P\{X_t = y_t \mid X_0^{t-1} = y_0^{t-1}\} = P\{X_t = y_t \mid X_{t-1} = y_{t-1}\}.$$

Toisin sanoin stokastisella prosessilla tarkoitetaan prosesseja, jotka etenevät ajan suhteen sattumanvaraisesti. Tässä tapauksessa Markovin prosessit alkavat joka hetki uudestaan, eivätkä muista menneisyyttään. Niiden tulevaisuuteen vaikuttaa vain nykytila, eikä se, miten siihen on päädytty.

Kaikille stokastisille prosesseille $\{X_t\}$ ja jokaiselle sekvenssille $y_0^t \in \mathbb{A}$ saadaan iteroimalla ehto

$$P\{X_0^t = y_0^t\} = P\{X_0 = y_0\} \prod_{i=0}^{t-1} P\{X_{i+1} = y_{i+1} \mid X_i = y_i\}.$$

Kaava todistaa lausekkeen $P\{X_{t+1} = j \mid X_t = i\}$ tärkeyden, missä näkyy kolme ominaisuutta: $i \in \mathbb{A}$ on "nykytila", $j \in \mathbb{A}$ on "seuraava tila"ja $t \in \mathbb{Z}$ on "nykyhetki". Merkitään

$$a_{ij} := P\{X_{t+1} = j \mid X_t = i\}.$$

Näin ollen a_{ij} on siirtymätodennäköisyys sille, kun ajan ollessa t siirrytään nykytilasta i seuraavaan tilaan j ja aikaan $t + 1$.

Määritelmä 4.4. Olkoon A stokastinen $n \times n$ matriisi, missä $A \in [0,1]^{n\times n}$, ja $\sum_{j=1}^{n} a_{ij} = 1$ kaikilla $i \in \mathbb{N}$. Tällöin vektorin $\pi \in S_n$ sanotaan olevan matriisin A *stationaarinen jakauma*, jos $\pi A = \pi$. Jos Markovin ketjulla on stationaarinen jakauma, niin kaikilla tulevilla tiloilla on sama jakauma.

4.2 Markovin piiloprosessin erimuotoiset mallit

Tässä osassa esitellään kolme erilaista Markovin piiloprosessin muotoa. Tämän jälkeen näytetään, että ne ovat keskenään ekvivalentit. Nämä kolme muotoa esiintyvät monissa kirjallisuuksissa ja tästä syystä on hyvä ymmärtää, että erilaisista ulkomuodoista riippumatta prosessit ovat keskenään ekvivalentit.

Määritelmä 4.5. Olkoon $\{Y_t\}_{t=1}^{\infty}$ stokastinen prosessi, jonka arvojoukko on $M = \{1, \ldots, m\}$. Sanotaan, että $\{Y_t\}$ on *tyypin 1 Markov piiloprosessi* tai *Markovin ketjun deterministisen funktion muoto*, jos on olemassa Markovin piiloprosessi $\{X_t\}_{t=0}^{\infty}$ yli äärellisen tilan $N = \{1, \ldots, n\}$ ja on olemassa funktio $f : N \to M$ siten että, $f(X_t) = Y_t$.

Määritelmä 4.6. Olkoon $\{Y_t\}_{t=1}^{\infty}$ stokastinen prosessi, jonka arvojoukko on $M = \{y_1, \ldots y_m\}$. Sanotaan, että $\{Y_t\}$ on *tyypin 2 Markovin piiloprosessi* tai *Markovin ketjun satunnaisfunktion muoto*, jos on olemssa luku n, matriisipari $A \in [0,1]^{n\times n}$, $B \in [0,1]^{n\times m}$ ja todennäköisyysjakauma $\pi \in S_n$ siten, että seuraavat ehdot täyttyvät

(1) A ja B ovat molemmat stokastisia matriiseja, joka tarkoittaa, että matriisien A ja B jokainen rivi summa on yksi tai ekvivalentisti

$$Ae_n = e_n, \quad Be_m = e_m,$$

missä $e_n = (1, 1, \ldots, 1)^\top$ ja $e_m = (1, 1, \ldots, 1)^\top$.

(2) Jakauma μ on A:n stationaarinen jakauma siten, että $\mu A = \mu$.

(3) Olkoon $\{X_t\}$ homogeeninen Markovin ketju tilan $N = \{1, \ldots, n\}$ mukaan, jolle on voimassa jakauma π ja siirtymämatriisi A. Täten

$$P\{X_0 = i\} = \pi_i \quad \text{ja} \quad P\{X_{t+1} = j \mid X_t = i\} = a_{ij}, \quad \forall i, j, t.$$

Olkoon satunnaismuuttuja Z_t jokaisella ajalla t

$$P\{Z_t = u \mid X_t = j\} = b_{ju}, \quad \forall j \in N,\ u \in M, t \geq 0.$$

Täten $\{Z_t\}$ on yhtä kuin $\{Y_t\}$.

Määritelmä 4.7. Olkoon $\{Y_t\}_{t=1}^{\infty}$ stokastinen prosessi, jonka arvojoukko on $M = \{y_1, \ldots, y_m\}$. Sanotaan, että $\{Y_t\}$ on *tyypin 3 Markovin prosessi* tai *yhdistetyn Markovin prosessin piiloprosessi*, jos on olemassa äärellinen joukko $N = \{1, \ldots, n\}$ ja stokastinen prosessi $\{X_t\}$ saa arvot joukosta N siten että, seuraavat ominaisuudet pätevät

1. Yhdistetty prosessi $\{(X_t, Y_t)\}$ on Markovin prosessi.

2. Lisäksi
$$P\{(X_t, Y_t) \mid (X_{t-1}, Y_{t-1})\} = P\{(X_t, Y_t) \mid X_{t-1}\},$$
joka voidaan esittää muodossa
$$P\{(X_t, Y_t) = (j, u) \mid (X_{t-1}, Y_{t-1}) = (i, v)\}$$
$$= P\{(X_t, Y_t) = (j, u) \mid X_{t-1} = i\}\ \forall i, j \in N\ ja\ \forall u, v \in M.$$

Tämän osion tarkoituksena on näyttää, että kaikki kolme HMP tyyppiä ovat samat niiden ulkomuodoistaan riippumatta. Kuitenkin, 3 tyypin Markovin piiloprosessi on kaikista yleisimmin käytetty, kun mallinnetaan ja lasketaan tiloja kun taas tyypin 1 Markovin piiloprosessi on vähiten yleinen. Tyypin 2 Markovin piiloprosessi on näiden kahden tyypin välissä.

Lause 4.8. *Seuraavat väitteet ovat ekvivalentit*

(i) Prosessilla $\{Y_t\}$ on tyypin 1 HMP ("Markovin ketjun deterministisen funktion" muoto).

(ii) Prosessilla $\{Y_t\}$ on tyypin 2 HMP ("Markovin ketjun satunnais funktion" muoto).

(iii) Prosessilla $\{Y_t\}$ on tyypin 3 HMP ("Yhdistetyn Markovin piiloprosessin" muoto).

Todistus. $(i) \Rightarrow (ii)$. Selvästi jokainen Markovin ketjun deterministinen funktio on myös saman Markovin ketjun "satunnais"funktio, kun B:n jokainen elementti on yhtä kuin nolla tai yksi. Koska molemmat joukot N ja M ovat äärellisiä, funktio f osittaa joukon N osiin $N_1, \ldots, N_m$, missä $N_u := \{i \in N : f(i) = u\}$. Täten joukossa N_u kaksi tilaa ovat mahdoton erottaa prosessissa $\{Y_t\}$. Nyt asetetaan $b_{ju} = 1$, jos $j \in N_u$ ja muuten nolla.

$(ii) \Rightarrow (iii)$. Jos $\{Y_t\}$ on mallinnettu tyypin 2 Markovin ketju, ja $\{X_t\}$ on siihen liitetty Markovin ketju, yhdistetty prosessi $\{(X_t, Y_t)\}$ on Markovin prosessi. Jos määritellään $(X_t, Y_t) \in N \times M$, niin seuraa HMP-ehdoista, että

$$P\{(X_{t+1}, Y_{t+1}) = (j, u) \mid (X_t, Y_t) = (i, v)\} = a_{ij} b_{ju},$$

ja on täten riippumaton muuttujasta v. Nyt määritellään

$$M^u := [a_{ij} b_{ju}] \in [0, 1]^{n \times n}.$$

Todennäköisyys, että $(X_{t+1}, Y_{t+1}) = (j, u)$ riippuu ainoastaan tilasta X_t eikä tarvitse ottaa huomioon tilaa Y_t. Tällöin yhdistetty prosessi $\{(X_t, Y_t)\}$ täyttää kaikki tyypin 3 HMP ehdot.

$(iii) \Rightarrow (i)$. Olkoon $\{Y_t\}$ tyypin 3 Markovin piiloprosessi ja olkoon Y_t Markovin prosessi siten että, yhdistetty prosessi $\{(X_t, Y_t)\}$ on myös Markovin prosessi. Nyt selvästi $f[(X_t, Y_t)] = Y_t$ sopivalle funktiolle f. Tällöin funktio on myös tyypin 1 Markovin piiloprosessi. □

Kuten mainittu aikaisemmin tyypin 3 Markovin prosessi on kaikista yleisin, ja se pätee myös sekvenssianalyysissä, jota esitellään seuraavassa osiossa. Vaikka sekvenssi esiintyy yhtenä kokonaisena ketjuna, voidaan sen katsoa osana pienempiä ketjuja, jolloin se on kokonaisuutena yhdistetty prosessi.

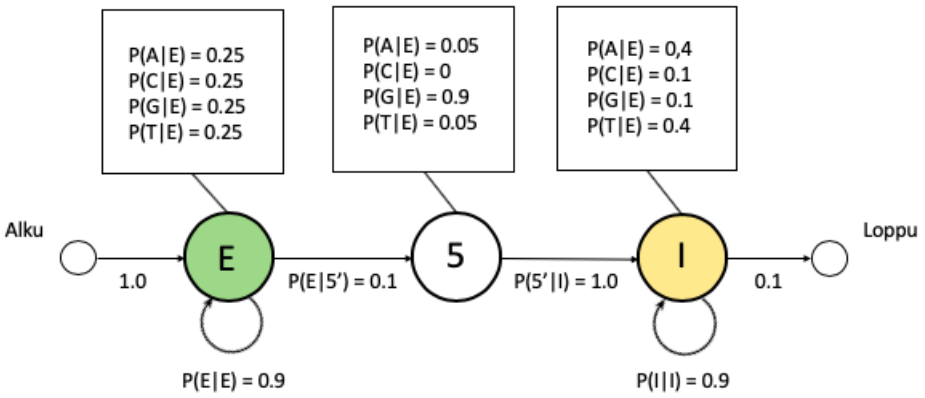
4.3 Markovin piiloprosessi sekvenssianalyysissä

On hyödyllistä ajatella HMP:n generoivan sekvenssiä. Jokaisella askeleella joko siirrytään tilasta toiseen tai pysytään samassa tilassa. Tila valitaan siirtymätodennäköi-

syysjakauman mukaan. Näin ollen prosessi generoi kaksi satunnaismuuttujamerkkijonoa: tilapolku (engl. *state path*), kun siirrytään yhdestä tilasta toiseen tilaan, ja havaintosekvenssi. Koska on annettu vain havaintosekvenssi, niin haluttu tilapolku on ”piilossa”. Tilapolkua kutsutaan *Markovin piiloketjuksi*.

Sekvenssianalyysissä tarkastellaan evaluointia ja estimointia. Evaluoinnissa halutaan tietää, että mikä on havaintosekvenssin todennäköisyys. Havaintosekvenssistä voidaan muodostaa potentiaalisesti monta tilapolkua, jotka voivat generoida saman sekvenssin. Halutaan löytää se tilapolku jolla on suurin todennäköisyys. Eli halutaan määrätään polku s, joka maksimoi edellisen todennäköisyyden $P(s \mid y_t)$. Tätä kutsutaan esimoinniksi.

Esimerkki 4.9. Esitellään yksinkertainen esimerkki 5’-pään tunnistamiselle. 5’-pää kuvaa nukleiinihapon viidettä hiiltä deoksiriboosissa, johon fosfaatti liittyy. Halutaan tunnistaa, missä vaiheessa tilan siirtyminen eksonista introniin tapahtuu — missä nukleotidissa on 5’-pää. Prosessissa jokainen nukleotidi vastaa yhtä seuraavista tiloista: E (eksoni), 5 (5’SS) tai I (introni). Tilat on järjestetty niin, että yhtä tai useampaa E tilaa seuraa yksi 5 tila ja tätä tilaa seuraa taas yksi tai useampi I tila. Näin ollen kuuden nukleotidin sekvenssin polku voisi olla AlkuEE5IIILoppu, mikä tarkoittaa, että kaksi ensimmäistä nukleotidia kuuluvat eksoniin ja kolme viimeistä nukleotidia introniin. Havainnollistetaan esimerkkiä alla olevalla kuvalla. Olkoon sekvenssi TTCGCTGACGTAACT. Nimetään mahdolliset tilapolut ja lasketaan niiden sekvenssien ilmentymistodennäköisyydet. Kuljetaan sekvenssi alusta loppuun ja tutkitaan, että mikä G-emäksisistä 5’SS paloista on todennäköisin?



Kuva 4.1. Markovin piiloprosessi

Kuvasta voidaan määrittää mahdolliset tilapolut:

$$s_1 = \text{AlkuEEE5IIIIIIIIIIILoppu}$$
$$s_2 = \text{AlkuEEEEEE5IIIIIIIILoppu}$$
$$s_3 = \text{AlkuEEEEEEEEE5IIIIILoppu}$$

Seuraavaksi estimoidaan suurin uskottavuus eli kerrotaan jokaisen tilapolun todennäköisyys yhteen. Todennäköisyys koostuu siirtymätodennäköisyydestä ja emissiotodennäköisyydestä (jokaisen tilan yläpuolella oleva taulukko). Kun sekvenssi käydään alusta loppuun, niin pitäisi saada 15 emissiota ja 16 siirtymää. Havaitaan laskemalla, että suurin 5'SS todennäköisyys sijaitsee kymmenennessä G nukleotidissa eli viimeisessä kolmannessa polussaz

$$P_{s_1} = P_{s_1,siirtymä} \cdot P_{s_1,emissio} = (1 \cdot 0.92 \cdot 0.1 \cdot 0.910 \cdot 0.1) \cdot (0.253 \cdot 0.95 \cdot 0.15 \cdot 0.46) \approx 1.72 \cdot 10^{-12}.$$
$$P_{s_2} = P_{s_2,siirtymä} \cdot P_{s_2,emissio} = (1 \cdot 0.95 \cdot 0.1 \cdot 0.97 \cdot 0.1) \cdot (0.256 \cdot 0.95 \cdot 0.13 \cdot 0.45) \approx 6.71 \cdot 10^{-12}.$$
$$P_{s_3} = P_{s_3,siirtymä} \cdot P_{s_3,emissio} = (1 \cdot 0.98 \cdot 0.1 \cdot 0.95 \cdot 0.1) \cdot (0.259 \cdot 0.95 \cdot 0.1 \cdot 0.44) \approx 2.62 \cdot 10^{-11}.$$

Monille ongelmille on useita mahdollisia tilapolkuja, joita ei pystytä käsin luottelemaan. Viterbin algortmi on dynaaminen ohjelmoitu algoritmi, joka on varma tapa löytää kaikista todennäköisin tilapolku.

Lähteet

[1] Happonen. P., Holopainen. M., Sariola. H., Sotkas. P., Tenhunen. A., Tihtarinen-Uimanen. M., Venäläinen. J. *Bios 5*. Sanoma pro Oy, Helsinki, 2014.

[2] Luosto. K. *Todennäköisyyslaskenta*. Tampereen Yliopisto, 2019.

[3] Nykter. M. *Introduction to Computational Biology*. Tampereen Yliopisto, Lecture, 5.9.2018.

[4] Vidyasagar. M. *Hidden Markov processes: theory and applications to biology*. Princeton University Press, 2014.