DNA-laskenta ja universaalisuus

Nisse Suutarinen
pro gradu matematiikassa
Turun yliopisto

2005

Sisältö

Johdanto

DNA-laskennalla tarkoitetaan algoritmista ongelmanratkaisua DNA-molekyylien ja niihin kohdistettavien reaktioiden avulla. Sen tutkimusta motivoivat DNA-molekyylien suotuisat ominaisuudet informaation tallennusmuotona ja molekyylien välisten reaktioiden samanaikaisuus, joka voisi mahdollistaa nopean rinnakkaislaskennan. Tutkimusalana DNA-laskenta on osa suuntausta, joka etsii luonnossa esiintyviä prosesseja hyödyntäviä laskentamalleja. Feynman [6] oli esittänyt jo vuonna 1960 idean biomolekyyleillä laskemisesta, mutta näyttöä bio- tai tarkemmin ottaen DNA-laskennan mahdollisesta käytännön toteuttavuudesta saatiin ensimmäisen kerran vuonna 1994, jolloin Adleman [1] suoritti kokeensa, jossa hän ratkaisi kauppamatkustajan ongelman erikoistapauksen koeputkessa DNA:ta ja laboratoriomenetelmiä käyttäen.

Koska olisi työlästä rakentaa jokaista ongelmaa kohti kyseisen ongelman ratkaiseva mekanismi, on tutkittu kysymystä, voidaanko rakentaa ohjelmoitava DNA-tietokone eli kiinteistä komponenteista koostuva DNA-molekyyleillä operointiin perustuva mekanismi, joka pystyy suorittamaan mitä tahansa annettua algoritmia. Teoreettisella tasolla on konstruoitu useita DNA-tietokoneen edellyttämiä Turingin koneiden laskentavoimalla varustettuja universaaleja malleja. Universaalisuus on laskentamallin ominaisuus suhteessa samantyyppisiin laskentamalleihin, joka tarkoittaa, että kyseisellä laskentamallilla voidaan simuloida mitä tahansa samantyyppistä laskentamallia.

Tässä työssä todistetaan kirjallisuutta käyttäen, että laajennetuiksi mH-systeemeiksi kutsuttu DNA-laskennan malli on laskentavoimaltaan ekvivalentti Turingin koneiden kanssa ja että on olemassa laajennettujen mH-systeemien suhteen universaali laajennettu mH-systeemi. Laajennetut mH-systeemit perustuvat silmukointioperaatioon, jonka loi Head [8] vuonna 1987 mallintamaan yhtä luonnon geneettisen materiaalin muuntelumekanismia, jossa restriktioentsyymien pilkkomat DNA-molekyylit yhdistyvät ristiin. Laajennetuiksi H-systeemeiksi kutsutun laajennettuja mH-systeemejä pelkistetymmän mallin osoitetaan saavuttavan korkeintaan äärellisten automaattien laskentavoiman. Lisäksi tarkastellaan komplementaarisuutta eli DNA:n nukleotidien pariutumista koskevaa sääntöä, joka luonnossa mahdollistaa perimän kahdentumisen, ja osoitetaan, että jonokoneiden eli tulostamiseen pystyvien äärellisten automaattien ja komplementaarisuuden avulla saavutetaan Turingin koneiden laskentavoima. Tällöin ei kuitenkaan päästä kiinteistä komponenteista koostuvaan malliin kuten laajennetuilla mH-systeemeillä. Toisaalta tällä tuloksella on yleistä mielenkiintoa, koska kaikki DNA-laskennan mallit hyödyntävät komplementaarisuutta jollakin tavalla ja komplementaarisuus saattaisi olla toteutettavissa jollakin muulla tavalla kuin DNA-molekyylien avulla. Esimerkiksi RNA-molekyyleillä on vastaava komplementaarisuusominaisuus.

Tarkastelut suoritetaan formaalisten kielten teorian puitteissa. Lukuun 1 on koottu luvuissa 2 ja 3 tarvittavat formaalisten kielten teorian käsitteet ja tulokset käyttäen lähteenä pääasiassa luentomonistetta [9] ja kirjoja [2], [10] ja [15]. Pykälässä 2.1 DNA-molekyylit abstrahoidaan sanoiksi ja määritellään komplementaarisuus morfismina käyttäen lähteenä kirjaa [12] ja artikkelia [17]. Pykälässä 2.2 tarkastellaan artikkelin [17] pohjalta biologisiksi palindromeiksi kutsuttuja sanoja, jotka liittyvät useiden restriktioentsyymien toimintaan. Pykälässä 2.3 tarkastellaan sekoitusoperaation ja komplementaarisuusmorfismilla varustetun aakkoston avulla määriteltävien TS- ja RTS-kielten ja DNA-molekyylien välistä yhteyttä käyttäen lähteenä kirjaa [12]. Pykälässä 2.1 osoitetaan kirjojen [16] ja [12] esitystä seuraten, että 0-tyypin kieliopeilla eli ekvivalentisti Turingin koneilla määriteltävissä olevat kielet voidaan esittää sekä TS- että RTS-kielen ja jonokoneiden avulla. Pykälässä 3.1 abstrahoidaan tietyllä tavalla yhteensopivien restriktioentsyymien pilkkomien DNA-molekyylien ristiinyhdistyminen sanojen väliseksi silmukointioperaatioksi. Pykälässä 3.2 määritellään silmukointioperaation avulla DNA-molekyylien joukoilla operoivia pykälässä 3.1 esitetyllä tavalla yhteensopivia restriktioentsyymejä mallintava laajennetuiksi H-systeemeiksi kutsuttu formaalin kielen generointimekanismi ja todistetaan artikkeliin [13] pohjautuen, että laajennettuja H-systeemejä voidaan simuloida äärellisillä automaateilla. Pykälässä 3.3 lisätään laajennettuihin H-systeemeihin multijoukkojen avulla kyky mallintaa molekyylejä, joita on tietty äärellinen määrä, ja todistetaan artikkelin [7] esitystä seuraten, että näin saaduilla laajennetuilla mH-systeemeillä voidaan simuloida mitä tahansa 0-tyypin kielioppia ja että on olemassa laajennettu mH-systeemi, jolla voidaan simuloida mitä tahansa laajennettua mH-systeemiä. Kysymyksessä on epäsuora simulointi, jossa tarvitaan Churchin-Turingin teesiä, kirjassa [10] todistettua tulosta, jonka mukaan Turingin koneita voidaan simuloida 0-tyypin kieliopeilla ja kirjassa [12] todistettua tulosta, jonka mukaan on olemassa universaali 0-tyypin kielioppi.

Biologisten termien valinnassa on käytetty sivustoa [3].

1 Formaalisten kielten teorian käsitteitä ja tuloksia

Tähän lukuun on koottu luvuissa 2 ja 3 tarvittavat formaalisten kielten teorian käsitteet ja tulokset sekä joitakin joukko-opin käsitteitä. Osa tuloksista todistetaan, kun taas osaan tuloksista on liitetty vain niiden todistuksissa tarvittavat konstruktiot tai viittaus kirjallisuudesta löytyvään todistukseen.

Tyhjästä joukosta käytetään merkintää $\emptyset$, joukon S potenssijoukosta eli joukon S osajoukkojen joukosta mukaanlukien tyhjä joukko käytetään merkintää $\mathcal{P}(S)$ ja joukon S alkioiden lukumäärästä merkintää $card(S)$. Luonnollisten lukujen joukon $\mathbb{N}$ oletetaan sisältävän nollan.

1.1 Multijoukot ja relaatiot

Multijoukko M yli joukon S on funktio

$$M : S \to \mathbb{N} \cup \{\infty\},$$

joka liittää jokaiseen joukon S alkioon x alkioiden x lukumäärän. Jos merkitään funktiota M samalla kirjaimella kuin joukkoa S, niin multijoukko S yli joukon S on funktio

$$S : S \to \mathbb{N} \cup \infty.$$

Multijoukon S yli joukon S alkiot, joita on vähintään yksi, muodostavat joukon S osajoukon

$$supp(S) = \{x \in S \mid S(x) > 0\}.$$

Jos $card(supp(S)) < \infty$, kysymyksessä on *äärellinen* multijoukko. Äärellinen multijoukko S yli joukon S voidaan esittää muodossa

$$\{(x, S(x)) \mid x \in supp(S)\}.$$

Jos M_1 ja M_2 ovat multijoukkoja yli joukon S, niin multijoukkojen M_1 ja M_2 *unioni* $M_1 \cup M_2$ on multijoukko yli joukon S, joka määräytyy ehdosta

$$(M_1 \cup M_2)(x) = M_1(x) + M_2(x) \quad \forall x \in S.$$

Jos multijoukot M_1 ja M_2 toteuttavat ehdon

$$M_1(x) \geq M_2(x) \quad \forall x \in S,$$

niin multijoukkojen M_1 ja M_2 *erotus* $M_1 - M_2$ on multijoukko yli joukon S, joka määräytyy ehdosta

$$(M_1 - M_2)(x) = M_1(x) - M_2(x) \quad \forall x \in S.$$

Joukon S *binäärirelaatio* R tai lyhyesti *relaatio* on joukon

$$S \times S = \{(a, b) \mid a, b \in S\}$$

osajoukko. Joukon S alkiot a ja b ovat *vertailtavia* relaation R suhteen, jos

$$(a, b) \in R \quad \text{tai} \quad (b, a) \in R.$$

Jos

$$(a, a) \in R \quad \text{aina, kun} \quad a \in S,$$

relaatio R on *refleksiivinen*. Jos

$$(a, c) \in R \quad \text{aina, kun} \quad (a, b) \in R \text{ ja } (b, c) \in R,$$

niin relaatio R on *transitiivinen*. Relaation R *refleksiivinen ja transitiivinen sulkeuma* on joukkojen sisältymisrelaation suhteen pienin joukon S relaatio, johon relaatio R sisältyy ja joka on refleksiivinen ja transitiivinen.

1.2 Sanat, kielet ja morfismit

Aakkosto on äärellinen epätyhjä joukko, jonka alkioita sanotaan *kirjaimiksi* tai *symboleiksi*. *Sana* yli aakkoston Σ on äärellinen jono aakkoston Σ kirjaimia, jossa sama kirjain voi toistua useamman kerran. Nollan kirjaimen jonoa sanotaan *tyhjäksi* sanaksi ja siitä käytetään merkintää λ. Kirjainten lukumäärästä sanassa w, kun kirjainten mahdolliset useampikertaiset esiintymät otetaan huomioon, eli sanan w *pituudesta* käytetään merkintää $|w|$ ja kirjainten a lukumäärästä sanassa w merkintää $|w|_a$. Aakkoston Σ yli muodostettujen sanojen joukosta mukaanlukien tyhjä sana käytetään merkintää Σ^* ja aakkoston Σ yli muodostettujen tyhjästä eroavien sanojen joukosta merkintää Σ^+. Tällöin

$$\Sigma^+ = \Sigma^* \backslash \{\lambda\}.$$

Jos u ja v ovat joukkoon Σ^* kuuluvia sanoja, niin sanat u ja v peräkkäin kirjoittamalla saatu sana uv kuuluu joukkoon Σ^*. Tätä operaatiota sanotaan *katenoinniksi* ja tyhjä sana on sen suhteen identiteettialkio eli

$$w\lambda = \lambda w = w \quad \forall w \in \Sigma^*.$$

Katenointioperaatio on lisäksi assosiatiivinen eli

$$u(vw) = (uv)w \quad \forall u, v, w \in \Sigma^*,$$

joten joukko Σ^* on katenoinnin suhteen monoidi. Sanan w katenointi itsensä kanssa voidaan esittää käyttäen potenssimerkintää

$$\begin{aligned} w^0 &= \lambda \\ w^n &= ww^{n-1} = w^{n-1}w, \end{aligned}$$

missä n on positiivinen kokonaisluku.

Katenoimalla tyhjästä eroavan sanan $w = a_1a_2 \cdots a_n$, missä $n \geq 1$ ja $a_i \in \Sigma$, kun $1 \leq i \leq n$, kirjaimet viimeisestä ensimmäiseen saadaan sanan w *peilikuva*

$$w^R = a_n a_{n-1} \cdots a_1.$$

Tyhjän sanan peilikuvaksi määritellään $\lambda^R = \lambda$. Ehdon

$$w^R = w$$

täyttäviä sanoja sanotaan *palindromeiksi*. Peilikuva- ja katenaatio-operaatioiden määritelmistä seuraa, että

$$(u^R)^R = u \quad \forall u \in \Sigma^* \tag{1}$$

$$(uv)^R = v^R u^R \quad \forall u, v \in \Sigma^*. \tag{2}$$

Sana v on sanan w *osasana*, jos on olemassa sellaiset sanat w_1 ja w_2, että

$$w = w_1 v w_2. \tag{3}$$

Sanan w osasanojen joukosta käytetään merkintää $F(w)$. Jos yhtälössä 3 $w_1 = \lambda$, niin sana v on sanan w *etuliite* ja jos $w_2 = \lambda$, sana v on sanan w *loppuliite*. Jos $v \neq w$, niin sanan w etu- tai loppuliite v on *aito*. Jos sana v on sanan w etuliite, niin merkitään $v \leq w$ ja jos sana v on sanan w aito etuliite, niin merkitään $v < w$. Jos $u \leq w$ eli $w = uv$, niin sanasta w voidaan *jakaa* sana u, jolloin merkitään $v = u^{-1}w$.

Joukon Σ^* äärellisiä tai äärettömiä osajoukkoja sanotaan *formaalisiksi kieliksi* tai lyhyesti *kieliksi* yli aakkoston Σ. Katenointioperaatio yleistyy kielten väliseksi operaatioksi siten, että jos L_1 ja L_2 ovat kieliä, niin

$$L_1 L_2 = \{uv \mid u \in L_1, v \in L_2\}.$$

Kielen L katenointi itsensä kanssa voidaan esittää käyttäen potenssimerkintää

$$\begin{aligned} L^0 &= \{\lambda\} \\ L^n &= LL^{n-1} = L^{n-1}L, \end{aligned}$$

missä n on positiivinen kokonaisluku. Kielen L *Kleenen* *- ja $+$-*sulkeumat* ovat

$$L^* = \bigcup_{i \geq 0} L^i \quad \text{ja} \quad L^+ = \bigcup_{i \geq 1} L^i.$$

Sanojen $x \in \Sigma_1^*$ ja $y \in \Sigma_2^*$ *sekoitus* on äärellinen kieli

$$x \sqcup y = \{x_1 y_1 x_2 y_2 \cdots x_n y_n \mid x = x_1 \cdots x_n, y = y_1 \cdots y_n, x_i \in \Sigma_1^*, y_i \in \Sigma_2^*, 1 \leq i \leq n\}.$$

Sekoitusoperaatio yleistyy luonnollisella tavalla kielille:

$$L_1 \sqcup L_2 = \bigcup_{w_1 \in L_1, w_2 \in L_2} w_1 \sqcup w_2.$$

Morfismi on kuvaus $h : \Sigma_1^* \to \Sigma_2^*$, missä Σ_1 ja Σ_2 ovat aakkostoja, joka toteuttaa ehdon

$$h(uv) = h(u)h(v) \quad \forall u, v \in \Sigma_1^*.$$

Se määräytyy täysin kuvien

$$h(a_i), \ a_i \in \Sigma_1$$

perusteella. Morfismin h *käänteismorfismi* on mahdollisesti osittain määritelty kuvaus $h^{-1} : \Sigma_2^* \to \mathcal{P}(\Sigma_1^*)$,

$$h^{-1}(w) = \{v \mid h(v) = w\}.$$

Projektiokuvaus aakkostolta Σ_1 aakkostolle Σ_2, missä $\Sigma_2 \subseteq \Sigma_1$, on morfismi $pr_{\Sigma_2} : \Sigma_1^* \to \Sigma_2^*$,

$$pr_{\Sigma_2}(a) = \begin{cases} a & \text{jos} \quad a \in \Sigma_2 \\ \lambda & \text{jos} \quad a \in \Sigma_1 \backslash \Sigma_2. \end{cases}$$

Morfismien $h_1, h_2 : \Sigma_1^* \to \Sigma_2^*$ *yhtäsuuruusjoukko* on kieli

$$E(h_1, h_2) = \{w \in \Sigma_1^+ \mid h_1(w) = h_2(w)\},$$

joka koostuu niistä tyhjästä eroavista sanoista, joilla morfismien h_1 ja h_2 kuvat yhtyvät.

Huomautus 1.1. Jos $h_1(w) = h_2(w)$, niin jompikumpi relaatioista

$$h_1(w') \leq h_2(w'), \quad h_2(w') \leq h_1(w')$$

toteutuu kaikilla sanan w etuliitteillä w'.

Huomautus 1.2. Tyhjä sana kuuluu kaikkien morfismien h määrittelyjoukkoon, ja koska

$$h(\lambda)h(\lambda) = h(\lambda\lambda) = h(\lambda),$$

kaikki morfismit kuvaavat tyhjän sanan tyhjäksi sanaksi. Ongelma, jossa kysytään, onko olemassa tyhjästä eroavaa sanaa, jolla annettujen morfismien kuvat yhtyvät, on osoitettu *ratkeamattomaksi*. Ei siis ole olemassa algoritmia, joka kaikilla annetuilla morfismeilla h_1, h_2 ratkaisisi, onko joukko $E(h_1, h_2)$ epätyhjä. Tuloksen todisti ensimmäisenä Post [14] vuonna 1946.

Esimerkki 1.3. Määritetään morfismien $f, g : \{a, b, c\}^* \to \{a, b\}^*$,

	a	b	c
f	a^2	b^2	ab^2
g	a^2b	ba	b,

yhtäsuuruusjoukko hyödyntäen huomautusta 1.1. Oletetaan, että morfismien f ja g yhtäsuuruusjoukossa on tyhjästä sanasta eroava sana w ja luetaan sitä kirjain kerrallaan vasemmalta oikealle. Koska sanat $f(b), g(b)$ ja $f(c), g(c)$ eivät ole vertailtavia relaation $\leq$ suhteen, sana w ei voi alkaa kirjaimilla b eikä c, joten sen täytyy alkaa kirjaimella a. Koska

$$\begin{array}{rclcrcl} f(aa) & = & a^4 & \quad & f(ac) & = & a^3b^2 \\ g(aa) & = & a^2ba^2b & & g(ac) & = & a^2b^2, \end{array}$$

sanan w täytyy alkaa sanalla ab ja koska

$$\begin{array}{rcl} f(ab^2) & = & a^2b^4 \\ g(ab^2) & = & a^2b^2aba, \end{array}$$

sana w ei voi alkaa sanalla ab^2. Laskelmista

$$\begin{array}{rclcrclcrcl} f(abca) & = & a^2b^2ab^2a^2 & \quad & f(abcb) & = & a^2b^2ab^4 & \quad & f(abcc) & = & a^2b^2ab^2ab^2 \\ g(abca) & = & a^2b^2aba^2b & & g(abcb) & = & a^2b^2ab^2a & & g(abcc) & = & a^2b^2ab^2 \end{array}$$

nähdään, että sana w ei voi alkaa sanalla abc, kun otetaan huomioon morfismien määrittely tapauksessa $abcc$. Laskelmista

$$\begin{array}{rclcrcl} f(abaa) & = & a^2b^2a^4 & \quad & f(abab) & = & a^2b^2a^2 \\ g(abaa) & = & a^2b^2a^3ba^2b & & g(abab) & = & a^2b^2a^3b^2a \end{array}$$

nähdään, että sanan w täytyy alkaa sanalla $abac$. Koska

$$\begin{array}{rcl} f(abac) & = & a^2b^2a^3b^2 \\ g(abac) & = & a^2b^2a^3b^2, \end{array} \tag{4}$$

ylläolevaa päättelyä toistamalla nähdään, että $w \in \{abac\}^+$, joten $E(f, g) \subseteq \{abac\}^+$. Toisaalta laskelmien 4 nojalla $\{abac\}^+ \subseteq E(f, g)$. Morfismien f ja g yhtäsuuruusjoukko on siis $\{abac\}^+$.

Esimerkki 1.4. Määritetään morfismien $f, g : \{a, b\}^* \to \{a, b\}^*$,

	a	b
f	a	baa
g	aab	aa,

yhtäsuuruusjoukko. Jos $f(w) = g(w)$, niin kirjainten a ja b lukumäärien sanoissa $f(w)$ ja $g(w)$ tulee vastata toisiaan eli

$$\begin{cases} f(w)_a &= g(w)_a \\ f(w)_b &= g(w)_b, \end{cases}$$

jolloin morfismien f ja g määrittelyn perusteella

$$\begin{cases} |w|_a + 2|w|_b &= 2|w|_a + 2|w|_b \\ |w|_b &= |w|_a. \end{cases}$$

Jälkimmäisestä yhtälöparista voidaan ratkaista $|w|_a = |w|_b = 0$, joten $w = \lambda$. Morfismien f ja g yhtäsuuruusjoukko on siis tyhjä.

On kuitenkin olemassa mielivaltaisen pitkiä sanoja w', jotka toteuttavat ehdon

$$f(w') \leq g(w'). \tag{5}$$

Osoitetaan induktiolla, että muotoa $a^2b^2a^4b^4a^8b^8 \cdots a^{2^n}b^{2^n}$ olevat sanat, missä n on positiivinen kokonaisluku, toteuttavat yhtälön

$$f(a^2b^2a^4b^4 \cdots a^{2^n}b^{2^n})g(b)^{2^n-1} = g(a^2b^2a^4b^4 \cdots a^{2^n}b^{2^n}) \tag{6}$$

ja siis myös ehdon 5. Kun $n = 1$,

$$f(a^2b^2)g(b) = a^2ba^2ba^4 = g(a^2b^2).$$

Oletetaan, että yhtälö 6 pätee, kun $n = k$. Tällöin

$$\begin{aligned} g(a^2b^2 \cdots a^{2^k}b^{2^k}a^{2^{k+1}}b^{2^{k+1}}) &= g(a^2b^2 \cdots a^{2^k}b^{2^k})g(a^{2^{k+1}}b^{2^{k+1}}) \\ &= f(a^2b^2 \cdots a^{2^k}b^{2^k})g(b)^{2^k-1}g(a^{2^{k+1}}b^{2^{k+1}}) \\ &= f(a^2b^2 \cdots a^{2^k}b^{2^k})(aa)^{2^k-1}(aab)^{2^{k+1}}(aa)^{2^{k+1}} \\ &= f(a^2b^2 \cdots a^{2^k}b^{2^k})a^{2^{k+1}-2}(aab)^{2^{k+1}}(aa)^{2^{k+1}} \\ &= f(a^2b^2 \cdots a^{2^k}b^{2^k})a^{2^{k+1}}(baa)^{2^{k+1}}(aa)^{2^{k+1}-1} \\ &= f(a^2b^2 \cdots a^{2^k}b^{2^k})f(a)^{2^{k+1}}f(b)^{2^{k+1}}g(b)^{2^{k+1}-1} \\ &= f(a^2b^2 \cdots a^{2^{k+1}}b^{2^{k+1}})g(b)^{2^{k+1}-1} \end{aligned}$$

eli yhtälö 6 pätee, kun $n = k + 1$, mikä todistaa väitteen.

1.3 Äärelliset automaatit ja jonokoneet

Äärelliset automaatit ovat kielten tunnistusmekanismeja, jotka joko *tunnistavat* eli *hyväksyvät* tai kielteisessä tapauksessa *hylkäävät* syötteenä saamansa sanan. Ne lukevat syötettä kirjain kirjaimelta ja edustavat sellaista algoritmisen laskettavuuden tasoa, joka on toteutettavissa syötteen vaatimaa talletustilaa lukuunottamatta puhtaasti äärellisin keinoin.

Yleistetyn äärellisen automaatin $M = (Q, \Sigma, \delta, Q_0, F)$ komponenteista Q on äärellinen *tilojen* joukko, Σ on *syöteaakkosto*,

$$\delta : Q \times (\Sigma \cup \{\lambda\}) \to \mathcal{P}(Q)$$

on *siirtymäfunktio*, $Q_0 \subseteq Q$ *alkutilojen* joukko ja $F \subseteq Q$ *lopputilojen* joukko. Joukon $Q \times \Sigma^*$ ehdosta

$$(q, ax) \vdash_M (q', x) \quad \Leftrightarrow \quad q' \in \delta(q, a),$$

missä $a \in \Sigma \cup \{\lambda\}$ ja $x \in \Sigma^*$, määräytyvän relaation $\vdash_M$ refleksiivisen ja transitiivisen sulkeuman $\vdash_M^*$ avulla voidaan määritellä yleistetyn äärellisen automaatin M *tunnistama* eli *hyväksymä* kieli

$$L(M) = \{x \in \Sigma^* \mid \exists q_0 \in Q_0, f \in F : (q_0, x) \vdash_M^* (f, \lambda)\}.$$

Se koostuu sellaisista sanoista x yli syöteaakkoston Σ, jotka luettuaan automaatti M voi olla jossakin lopputilassa, kun lukeminen aloitetaan jostakin alkutilasta. Yleistetty äärellinen automaatti on siis epädeterministinen malli.

Yleistetyt äärelliset automaatit M ja M' ovat *ekvivalentit*, jos

$$L(M) = L(M').$$

Yleistettyjen äärellisten automaattien tunnistamien kielten luokkaa sanotaan *säännöllisten* kielten luokaksi ja siitä käytetään merkintää *Reg*.

Yleistetty äärellinen automaatti $M = (Q, \Sigma, \delta, Q_0, F)$ voidaan esittää *graafina* muodossa

$$((Q, E, \epsilon), Q_0, F),$$

missä Q_0 ja F ovat automaatin M alku- ja lopputilojen joukot ja (Q, E, ϵ) on suunnattu leimattu graafi. Graafin (Q, E, ϵ) pisteet ovat automaatin M tilat Q ja sen pisteiden välillä on viivat

$$E = \{(q, q') \in Q \times Q \mid q' \in \delta(q, a) \text{ joillakin } q, q' \in Q, a \in \Sigma \cup \{\lambda\}\},$$

jotka on leimattu sanoilla

$$\epsilon((q, q')) = \;\; a \quad \text{jos} \quad q' \in \delta(q, a),$$

missä ϵ on morfismi $E^* \to \Sigma^*$. Kääntäen jokaista edelläkuvattua muotoa olevaa esitystä $((Q, E, \epsilon), Q_0, F)$ vastaa yksikäsitteinen yleistetty äärellinen automaatti. Graafin (Q, E, ϵ) polku on sana

$$e_1 e_2 \cdots e_n \in E^*, \quad n \geq 0,$$

jonka kirjaimet $e_i \in E$ ovat muotoa

$$e_i = (q_i, q_{i+1}), \quad i = 1, 2, \ldots, n.$$

Tyhjästä eroava polku on *hyväksyvä*, jos $q_1 \in Q_0$ ja $q_{n+1} \in F$. Tyhjä polku on hyväksyvä, jos $Q_0 \cap F \neq \emptyset$. Koska leimafunktio ϵ on morfismi $E^* \to \Sigma^*$, se on määritelty kaikilla graafin (Q, E, ϵ) poluilla ja erityisesti kaikilla hyväksyvillä poluilla.

Deterministisen äärellisen automaatin $M = (Q, \Sigma, \delta, \{q_0\}, F)$ komponentit ovat muutoin samat kuin yleistetyllä äärellisellä automaatilla, mutta sillä on vain yksi alkutila ja sen siirtymäfunktio

$$\delta : Q \times \Sigma \to \mathcal{P}(Q),$$

joka ei salli muotoa $\delta(q, \lambda)$ olevia siirtymiä, toteuttaa ehdon

$$card(\delta(q, a)) \leq 1 \quad \forall q \in Q, a \in \Sigma. \tag{7}$$

Deterministisen äärellisen automaatin siirtymäfunktio voidaan määritellä myös osittain määriteltynä funktiona

$$\delta : Q \times \Sigma \to Q.$$

Tällöin $\delta(p, a)$ ei ole määritelty niissä tapauksissa, joissa

$$\delta(p, a) = \emptyset$$

käyttäen siirtymäfunktion maalijoukkona joukkoa $\mathcal{P}(Q)$.

Determinististä äärellistä automaattia voidaan pitää yleistetyn äärellisen automaatin erikoistapauksena, joten sen tunnistama kieli määritellään vastaavasti kuin yleistetyillä äärellisillä automaateilla.

Lause 1.5. *Jokaista säännöllistä kieltä L kohti on olemassa deterministinen äärellinen automaatti, jonka tunnistama kieli on L.*

Johdetaan todistuksessa tarvittava konstruktio käyttäen lähteenä luentomonistetta [9]. Olkoon $M_1 = (Q, \Sigma, \delta_1, Q_0, F)$ yleistetty äärellinen automaatti, jonka tunnistama kieli on L. Yleistetty äärellinen automaatti

$$M_2 = (Q \cup \{q_0\}, \Sigma, \delta_2, \{q_0\}, F),$$

missä $q_0 \notin Q$, jolla on kaikki automaatin M_1 siirtymät eli

$$q \in \delta_1(p, a) \quad \Rightarrow \quad q \in \delta_2(p, a)$$

aina, kun $p \in Q$ ja $a \in \Sigma \cup \{\lambda\}$ ja jolla on lisäksi siirtymät

$$\delta_2(q_0, a) = \{q \mid \exists p \in Q_0 : q \in \delta_1(p, a)\},$$

missä $a \in \Sigma \cup \{\lambda\}$, on automaatin M_1 kanssa ekvivalentti ja sillä on vain yksi alkutila.

Määritellään jokaista automaatin M_2 tilaa $p \in Q \cup \{q_0\}$ kohti joukot

$$\begin{aligned} C_0(p) &= \{p\}, \\ C_{i+1}(p) &= C_i(p) \cup \{q \in Q \cup \{q_0\} \mid \exists r \in C_i(p) : q \in \delta_2(r, \lambda)\}, i = 0, 1, 2, \ldots. \end{aligned}$$

Koska $C_i(p) \subseteq C_{i+1}(p)$, lukujono

$$card(C_0(p)), card(C_1(p)), card(C_2(p)), \ldots$$

on kasvava. Lisäksi luku $card(Q) + 1$ rajoittaa sen jäseniä ylhäältä. On siis olemassa sellainen indeksi k, että $C_{k+1}(p) = C_k(p)$. Olkoon i_p pienin tämän ehdon toteuttava indeksi. Yleistetty äärellinen automaatti

$$M_3 = (Q \cup \{q_0\}, \Sigma, \delta_3, C_{i_{q_0}}(q_0), F),$$

jolla on siirtymät

$$q \in \delta_3(p, a) \quad \Leftrightarrow \quad \exists r \in Q \cup \{q_0\} : r \in \delta_2(p, a) \text{ ja } q \in C_{i_r}(r)$$

kaikilla $p, q \in Q \cup \{q_0\}, a \in \Sigma$, ei sisällä muotoa $\delta_3(p, \lambda)$ olevia siirtymiä ja toteuttaa $L(M_3) = L(M_2)$.

Deterministiselle äärelliselle automaatille

$$M = (\mathcal{P}(Q \cup \{q_0\}), \Sigma, \delta, \{C_{i_{q_0}}(q_0)\}, \{H \in \mathcal{P}(Q \cup \{q_0\}) \mid H \cap F \neq \emptyset\}),$$

jolla on siirtymät

$$\delta(H, a) = \{q \in Q \cup \{q_0\} \mid \exists p \in H : q \in \delta_3(p, a)\}$$

kaikilla $H \in \mathcal{P}(Q \cup \{q_0\}), a \in \Sigma$, pätee

$$L(M) = L(M_3) = L(M_2) = L(M_1) = L.$$

Lause 1.6. *Säännöllisten kielten luokka on suljettu unionin, katenoinnin, sekoituksen, Kleenen *- ja +-sulkeumien, morfisen kuvan ja leikkauksen suhteen.*

Esitetään lauseen todistuksessa tarvittavat konstruktiot käyttäen lähteenä luentomonistetta [9] ja sekoitusoperaation osalta kirjaa [5]. Olkoot $L_1, L_1' \subseteq \Sigma_1^*$ ja $L_2 \subseteq \Sigma_2^*$ säännöllisiä kieliä, $h : \Sigma_1^* \to \Sigma_2^*$ morfismi ja

$$\begin{aligned} M_i &= (Q_i, \Sigma_i, \delta_i, Q_{0,i}, F_i),\ i = 1, 2 \\ M_1' &= (Q_1', \Sigma_1', \delta_1', Q_{0,1}', F_1') \end{aligned}$$

yleistettyjä äärellisiä automaatteja, joiden tilojen joukot ovat erilliset ja joiden tunnistamat kielet ovat $L(M_1) = L_1$, $L(M_2) = L_2$ ja $L(M_1') = L_1'$.

Yleistetty äärellinen automaatti

$$M_3 = (Q_1 \cup Q_2, \Sigma_1 \cup \Sigma_2, \delta_3, Q_{0,1} \cup Q_{0,2}, F_1 \cup F_2),$$

jolla on siirtymät

$$q \in \delta_3(p, a) \quad \Leftrightarrow \quad q \in \delta_1(p, a) \text{ tai } q \in \delta_2(p, a)$$

kaikilla $p, q \in Q_1 \cup Q_2, a \in \Sigma_1 \cup \Sigma_2 \cup \{\lambda\}$, toteuttaa $L(M_3) = L_1 \cup L_2$.

Yleistetty äärellinen automaatti

$$M_4 = (Q_1 \cup Q_2, \Sigma_1 \cup \Sigma_2, \delta_4, Q_{0,1}, F_2),$$

jolla on siirtymät

$$q \in \delta_1(p, a) \text{ tai } q \in \delta_2(p, a) \quad \Rightarrow \quad q \in \delta_4(p, a)$$

kaikilla $p, q \in Q_1 \cup Q_2, a \in \Sigma_1 \cup \Sigma_2 \cup \{\lambda\}$ ja lisäksi siirtymät

$$Q_{0,2} \subseteq \delta_4(f, \lambda),$$

kaikilla $f \in F_1$, toteuttaa $L(M_4) = L_1 L_2$.

Yleistetty äärellinen automaatti

$$M_5 = (Q_1 \times Q_2, \Sigma_1 \cup \Sigma_2, \delta_5, Q_{0,1} \times Q_{0,2}, F_1 \times F_2),$$

jolla on siirtymät

$$\delta((q_1, q_2), a) = \{(q_1', q_2) \mid q_1' \in \delta_1(q_1, a)\} \cup \{(q_1, q_2') \mid q_2' \in \delta_2(q_2, a)\}$$

kaikilla $(q_1, q_2) \in Q_1 \times Q_2, a \in (\Sigma_1 \cup \Sigma_2 \cup \{\lambda\})$, toteuttaa $L(M_5) = L_1 \sqcup\!\sqcup L_2$.

Yleistetty äärellinen automaatti

$$M_6 = (Q_1 \cup \{q'\}, \Sigma_1, \delta_6, \{q'\}, F_1 \cup \{q'\}),$$

missä $q' \notin Q_1$, jolla on kaikki automaatin M_1 siirtymät eli

$$q \in \delta_1(p, a) \Rightarrow q \in \delta_6(p, a)$$

kaikilla $p, q \in Q_1, a \in \Sigma_1 \cup \{\lambda\}$ ja jolla on lisäksi siirtymät

$$\delta_6(q', \lambda) = Q_{0,1} \quad \text{ja} \quad q' \in \delta_6(f, \lambda),$$

kaikilla $f \in F_1$, toteuttaa $L(M_6) = L_1^*$.

Yleistetty äärellinen automaatti

$$M_7 = (Q_1, \Sigma_1, \delta_7, Q_{0,1}, F_1),$$

jolla on kaikki automaatin M_1 siirtymät ja jolla on lisäksi siirtymät

$$Q_{0,1} \subseteq \delta_7(f, \lambda)$$

kaikilla $f \in F_1$, toteuttaa $L(M_7) = L_1^+$.

Lauseen 1.5 nojalla on olemassa deterministinen äärellinen automaatti $M_8 = (Q, \Sigma_1, \delta_8, \{q_0\}, F)$, jonka tunnistama kieli on L_1. Muodostetaan jokaista ehdon

$$|h(a)| = |a_1 a_2 \cdots a_n| > 1,$$

missä $a_i \in \Sigma_2$, kun $1 \leq i \leq n$, toteuttavaa kirjainta $a \in \Sigma_1$ kohti joukko

$$Q_a = \{q_{a,1}, q_{a,2}, \ldots, q_{a,n-1}\}$$

siten, että $Q_a \cap Q_b = \emptyset$ aina, kun $a \neq b$. Yleistetty äärellinen automaatti

$$M_9 = (Q \cup \bigcup_{a \in \Sigma_1, |h(a)| > 1} Q_a, \Sigma_2, \delta_9, \{q_0\}, F),$$

jolla on tapauksessa

$$|h(a)| = |a_1 a_2 \cdots a_n| > 1, \quad a \in \Sigma_1$$

siirtymät

$$\begin{aligned}
\delta_9(p, a_1) &= \{q_{a,1}\} \\
\delta_9(q_{a,1}, a_2) &= \{q_{a,2}\} \\
\delta_9(q_{a,2}, a_3) &= \{q_{a,3}\} \\
&\vdots \\
\delta_9(q_{a,n-2}, a_{n-1}) &= \{q_{a,n-1}\} \\
\delta_9(q_{a,n-1}, a_n) &= \{q\},
\end{aligned}$$

missä $q \in \delta_8(p, a)$ ja jolla on tapauksessa

$$|h(a)| \leq 1, \quad a \in \Sigma_1$$

siirtymät

$$q \in \delta_9(p, h(a)) \quad \Leftrightarrow \quad q \in \delta_8(p, a),$$

toteuttaa $L(M_9) = h(L_1)$.

Yleistetty äärellinen automaatti

$$M_{10} = (Q_1 \times Q'_1, \Sigma_1, \delta_{10}, Q_{0,1} \times Q'_{0,1}, F_1 \times F'_1),$$

jolla on siirtymät

$$\delta_{10}((p_1, p_2), a) = \{(q_1, q_2) \mid q_1 \in \delta_1(p_1, a), q_2 \in \delta'_1(p_2, a)\}$$

kaikilla $(p_1, p_2) \in Q_1 \times Q'_1, a \in \Sigma_1 \cup \{\lambda\}$, toteuttaa $L(M_{10}) = L_1 \cap L'_1$.

Lemma 1.7. *Jokainen äärellinen kieli on säännöllinen.*

Todistus. Olkoon $\Sigma = \{a_1, a_2, \ldots, a_n\}$ aakkosto ja $L \subseteq \Sigma^*$ äärellinen kieli. Deterministiset äärelliset automaatit

$$M_i = (\{q_{0,i}, q_{1,i}\}, \{a_i\}, \delta_i, \{q_{0,i}\}, \{q_{1,i}\}), i = 1, 2, \ldots, n,$$

joilla on siirtymät

$$\delta_i(q_{0,i}, a_i) = q_{1,i}, i = 1, 2, \ldots, n$$

toteuttavat $L(M_i) = \{a_i\}$.

Jokainen kielen L sana kuuluu kieleen, joka on kielten $\{a_i\}$ äärellinen katenaatio. Kielen L äärellisyyden nojalla kieli L voidaan siis ilmaista kielten $\{a_i\}$ äärellisen monen äärellisen katenaation unionina, joten lauseen 1.6 nojalla on olemassa yleistetty äärellinen automaatti, jonka tunnistama kieli on L. □

Jonokone $M = (Q, \Sigma, \Delta, \delta, \gamma, \{q_0\}, F)$ on deterministinen äärellinen automaatti, joka voi kunkin siirtymänsä yhteydessä tulostaa sanan. Sen komponenteista Q, Σ, $\{q_0\}$ ja F ovat kuten deterministisellä äärellisellä automaatilla, Δ on *tulosteaakkosto*,

$$\delta : Q \times \Sigma \to Q$$

mahdollisesti osittain määritelty siirtymäfunktio ja

$$\gamma : Q \times \Sigma \to \Delta^*$$

on *tulostefunktio*, jolla on sama määrittelyjoukko kuin siirtymäfunktiolla. Funktioiden δ ja γ määrittelyjoukot voidaan laajentaa joukoksi $Q \times \Sigma^*$ määrittelemällä

$$\begin{aligned}
\delta(q, \lambda) &= q & (8)\\
\gamma(q, \lambda) &= \lambda & (9)\\
\delta(q, fx) &= \delta(\delta(q, f), x) & (10)\\
\gamma(q, fx) &= \gamma(q, f)\gamma(\delta(q, f), x), & (11)
\end{aligned}$$

missä $f \in \Sigma^*$ ja $x \in \Sigma$.

Seuraavan lemman ja sitä seuraavan lauseen 1.9 todistukset ovat peräisin kirjasta [2].

Lemma 1.8. *Syöteaakkostolla Σ varustetun jonokoneen siirtymä- ja tulostefunktioille δ ja γ pätee*

$$\begin{aligned}
\delta(q, fg) &= \delta(\delta(q, f), g) \quad \forall f, g \in \Sigma^* & (12)\\
\gamma(q, fg) &= \gamma(q, f)\gamma(\delta(q, f), g) \quad \forall f, g \in \Sigma^*. & (13)
\end{aligned}$$

Todistus. Todistetaan yhtälöt 12 ja 13 induktiolla sanan g pituuden suhteen. Jos $|g| = 0$ eli $g = \lambda$, kummatkin yhtälöistä 12 ja 13 pätevät määritelmien 8 ja 9 nojalla. Jos $|g| \geq 1$ ja $g = hx$, missä $h \in \Sigma^*$ ja $x \in \Sigma$, ja oletetaan, että yhtälö 12 pätee sanaa g lyhyemmillä sanoilla, niin määritelmän 10 ja induktio-oletuksen nojalla

$$\begin{aligned}
\delta(q, fg) &= \delta(q, fhx)\\
&= \delta(\delta(q, fh), x)\\
&= \delta(\delta(\delta(q, f), h), x)\\
&= \delta(\delta(q, f), hx)\\
&= \delta(\delta(q, f), g),
\end{aligned}$$

joten yhtälö 12 pätee kaikilla sanoilla g. Oletetaan, että yhtälö 13 pätee sanaa g lyhyemmillä sanoilla, jolloin määritelmän 11, induktio-oletuksen ja yhtälön 12 nojalla

$$\begin{aligned}
\gamma(q, fg) &= \gamma(q, fhx)\\
&= \gamma(q, fh)\gamma(\delta(q, fh), x)\\
&= \gamma(q, f)\gamma(\delta(q, f), h)\gamma(\delta(\delta(q, f), h), x)\\
&= \gamma(q, f)\gamma(\delta(q, f), hx)\\
&= \gamma(q, f)\gamma(\delta(q, f), g),
\end{aligned}$$

joten yhtälö 13 pätee kaikilla sanoilla g. $\square$

Jos $M = (Q, \Sigma, \Delta, \delta, \gamma, \{q_0\}, F)$ on jonokone, niin mahdollisesti osittain määriteltyä kuvausta $g : \Sigma^* \to \Delta^*$,

$$g(u) = v \quad \Leftrightarrow \quad \begin{cases} \gamma(q_0, u) & = & v \\ \delta(q_0, u) & \in & F \end{cases}$$

sanotaan jonokoneen M määrittelemäksi gsm-kuvaukseksi. gsm-kuvaukset ovat syötteen ja tulosteen vaatimaa talletustilaa lukuunottamatta puhtaasti äärellisin keinoin laskettavissa olevia funktioita.

Lause 1.9. *Jos $f' : \Sigma^* \to \Delta^*$ ja $g' : \Delta^* \to \Gamma^*$ ovat gsm-kuvauksia, niin on olemassa gsm-kuvaus $h' : \Sigma^* \to \Gamma^*$, joka on kuvausten g' ja f' kompositio.*

Todistus. Olkoot gsm-kuvauksia f' ja g' vastaavat jonokoneet

$$\begin{aligned} M_1 &= (Q_1, \Sigma, \Delta, \delta_1, \gamma_1, \{q_{0,1}\}, F_1) \\ M_2 &= (Q_2, \Delta, \Gamma, \delta_2, \gamma_2, \{q_{0,2}\}, F_2). \end{aligned}$$

Määritellään jonokone

$$M_3 = (Q_2 \times Q_1, \Sigma, \Gamma, \delta_3, \gamma_3, \{(q_{0,2}, q_{0,1})\}, F_2 \times F_1),$$

jolla on siirtymät ja tulosteet

$$\delta_3((p, q), a) = (\delta_2(p, \gamma_1(q, a)), \delta_1(q, a)) \tag{14}$$
$$\gamma_3((p, q), a) = \gamma_2(p, \gamma_1(q, a)) \tag{15}$$

kaikilla $(p, q) \in Q_2 \times Q_1, a \in \Sigma$. Laajennetaan jonokoneiden M_1, M_2 ja M_3 siirtymä- ja tulostefunktioiden määrittelyjoukoiksi $Q_1 \times \Sigma^*$, $Q_2 \times \Delta^*$ ja $(Q_2 \times Q_1) \times \Sigma^*$ yhtälöjen 8, 9, 10 ja 11 osoittamalla tavalla. Osoitetaan induktiolla sanan f pituuden suhteen, että funktiot δ_3 ja γ_3 toteuttavat kaikilla $f \in \Sigma^*, (p, q) \in Q_2 \times Q_1$ yhtälöt

$$\delta_3((p, q), f) = (\delta_2(p, \gamma_1(q, f)), \delta_1(q, f)) \tag{16}$$
$$\gamma_3((p, q), f)) = \gamma_2(p, \gamma_1(q, f)). \tag{17}$$

Jos $|f| = 0$, yhtälön 16 oikea puoli saa yhtälöjen 8 ja 9 nojalla arvon (p, q), joka on sama kuin yhtälön vasen puoli yhtälön 8 nojalla. Oletetaan, että $f = hx$, missä $h \in \Sigma^*$ ja $x \in \Sigma$, ja otetaan selvyyden vuoksi käyttöön merkinnät $w = \gamma_1(q, h)$ ja $w' = \gamma_1(\delta_1(q, h), x)$, jolloin yhtälön 11 nojalla

$$\begin{aligned} ww' &= \gamma_1(q, h)\gamma_1(\delta_1(q, h), x) \\ &= \gamma_1(q, hx) \\ &= \gamma_1(q, f). \end{aligned} \tag{18}$$

Jos yhtälö 16 pätee sanaa f lyhyemmillä sanoilla, niin yhtälön 10, induktio-oletuksen, määritelmän 14 ja yhtälön 12 nojalla

$$\begin{aligned}
\delta_3((p,q),f) &= \delta_3((p,q),hx) \\
&= \delta_3(\delta_3((p,q),h),x) \\
&= \delta_3((\delta_2(p,\gamma_1(q,h)),\delta_1(q,h)),x) \\
&= (\delta_2(\delta_2(p,\gamma_1(q,h)),\gamma_1(\delta_1(q,h),x)),\delta_1(\delta_1(q,h),x)) \\
&= (\delta_2(\delta_2(p,w),w'),\delta_1(\delta_1(q,h),x)) \\
&= (\delta_2(p,ww'),\delta_1(q,hx)) \\
&= (\delta_2(p,\gamma_1(q,f),\delta_1(q,f)), \qquad (19)
\end{aligned}$$

joten yhtälö 16 pätee kaikilla sanoilla f.

Jos $f=\lambda$, yhtälön 17 oikea puoli saa yhtälön 9 nojalla arvon λ, joka on sama kuin yhtälön vasen puoli yhtälön 9 nojalla. Jos yhtälö 17 pätee sanaa f lyhyemmillä sanoilla, niin käyttäen samoja merkintöjä kuin yllä ja yhtälöä 11, induktio-oletusta, yhtälöä 16, määritelmää 15 ja yhtälöä 13

$$\begin{aligned}
\gamma_3((p,q),f) &= \gamma_3((p,q),hx) \\
&= \gamma_3((p,q),h)\gamma_3(\delta_3((p,q),h),x) \\
&= \gamma_2(p,\gamma_1(q,h))\gamma_3((\delta_2(p,\gamma_1(q,h)),\delta_1(q,h)),x) \\
&= \gamma_2(p,w)\gamma_3((\delta_2(p,w),\delta_1(q,h)),x) \\
&= \gamma_2(p,w)\gamma_2(\delta_2(p,w),\gamma_1(\delta_1(q,h),x)) \\
&= \gamma_2(p,w)\gamma_2(\delta_2(p,w),w') \\
&= \gamma_2(p,ww') \\
&= \gamma_2(p,\gamma_1(q,f)),
\end{aligned}$$

joten yhtälö 17 pätee kaikilla sanoilla f.

Olkoon jonokoneen M_3 määrittelemä gsm-kuvaus $h':\Sigma^*\to\Gamma^*$. Jos $w\in\Sigma^*$ on sellainen sana, että sanat $f'(w)\in\Delta^*$ ja $g'(f'(w))\in\Gamma^*$ on määritelty eli

$$\begin{aligned}
\delta_1(q_{0,1},w) &\in F_1 \\
\gamma_1(q_{0,1},w) &= f'(w)
\end{aligned}
\quad \text{ja} \quad
\begin{aligned}
\delta_2(q_{0,2},f'(w)) &\in F_2 \\
\gamma_2(q_{0,2},f'(w)) &= g'(f'(w)),
\end{aligned}$$

niin yhtälön 16 nojalla

$$\begin{aligned}
\delta_3((q_{0,2},q_{0,1}),w) &= (\delta_2(q_{0,2},\gamma_1(q_{0,1},w)),\delta_1(q_{0,1},w)) \\
&= (\delta_2(q_{0,2},f'(w)),\delta_1(q_{0,1},w)) \\
&\in F_2\times F_1
\end{aligned}$$

ja yhtälön 17 nojalla

$$\begin{aligned}\gamma_3((q_{0,2}, q_{0,1}), w) &= \gamma_2(q_{0,2}, \gamma_1(q_{0,1}, w))\\ &= \gamma_2(q_{0,2}, f'(w))\\ &= g'(f'(w)),\end{aligned}$$

joten jonokone M_3 määrittelee gsm-kuvauksen $g' \circ f'$. □

Lemma 1.10. *Jokainen morfismi on gsm-kuvaus.*

Todistus. Jos $f : \{a_1, \ldots, a_n\}^* \to \{b_1, \ldots, b_m\}^*$ on morfismi, niin yksitilaisen jonokoneen

$$M = (\{q\}, \{a_1, \ldots, a_n\}, \{b_1, \ldots, b_m\}, \delta, \gamma, \{q\}, \{q\}),$$

jolla on siirtymät

$$\delta(q, a_i) = q, \quad i = 1, 2, \ldots, n$$

ja tulosteet

$$\gamma(q, a_i) = f(a_i), \quad i = 1, 2, \ldots, n,$$

määrittelemä gsm-kuvaus g toteuttaa

$$f(w) = g(w) \quad \forall w \in \{a_1, \ldots, a_n\}^*.$$

□

Esimerkki 1.11. Olkoon $h : \{a_1, a_2, \ldots, a_n\}^* \to \{0, 1\}^*$ morfismi

$$h(a_i) = 01^i0, \quad i = 1, 2, \ldots, n.$$

Jonokoneen $M = (\{q\}, \{a_1, a_2, \ldots, a_n\}, \{0, 1\}, \delta, \gamma, \{q\}, \{q\})$, jolla on siirtymät

$$\delta(q, a_i) = q, \quad i = 1, 2, \ldots, n$$

ja tulosteet

$$\gamma(q, a_i) = 01^i0, \quad i = 1, 2, \ldots, n,$$

määrittelemä gsm-kuvaus $g : \{a_1, a_2, \ldots, a_n\}^* \to \{0, 1\}^*$ toteuttaa ehdon

$$g(w) = h(w) \quad \forall w \in \{a_1, a_2, \ldots, a_n\}^*.$$

Esimerkki 1.12. Olkoon $pr_{\Sigma_2} : \Sigma_1^* \to \Sigma_2^*$ projektiokuvaus aakkostolta Σ_1 aakkostolle Σ_2, missä $\Sigma_2 \subseteq \Sigma_1$. Jonokoneen

$$M = (\{q\}, \Sigma_1, \Sigma_2, \delta, \gamma, \{q\}, \{q\}),$$

jolla on siirtymät

$$\delta(q,a) = q \quad \forall a \in \Sigma_1$$

ja tulosteet

$$\gamma(q,a) = \begin{cases} \lambda & \text{jos} \quad a \in \Sigma_1 \backslash \Sigma_2 \\ a & \text{jos} \quad a \in \Sigma_2, \end{cases}$$

määrittelemä *gsm*-kuvaus $g : \Sigma_1^* \to \Sigma_2^*$ toteuttaa ehdon

$$g(w) = pr_{\Sigma_2}(w) \quad \forall w \in \Sigma_1^*.$$

Esimerkki 1.13. Olkoon $L \subseteq \Sigma^*$ mielivaltainen kieli ja $L' \subseteq \Sigma^*$ säännöllinen kieli. Lauseen 1.5 nojalla on olemassa deterministinen äärellinen automaatti $M = (Q, \Sigma, \delta, \{q_0\}, F)$, jonka tunnistama kieli on L'. Jonokone

$$M' = (Q, \Sigma, \Sigma, \delta, \gamma, \{q_0\}, F),$$

jonka syöte- ja tulosteaakkosto on Σ ja jolla on tulosteet

$$\gamma(q,a) = a$$

kaikilla pareilla (q, a), joilla tulostefunktio on määritelty, määrittelee *gsm*-kuvauksen $g : \Sigma^* \to \Sigma^*$,

$$g(u) = u \quad \Leftrightarrow \quad \delta(q_0, u) \in F,$$

jonka määrittelyjoukko on L'. Tällöin

$$\begin{aligned} g(L) &= \{g(w) \mid w \in L\} \\ &= \{g(w) \mid w \in L \cap L'\} \\ &= \{w \mid w \in L \cap L'\} \\ &= L \cap L'. \end{aligned}$$

1.4 0-tyypin kieliopit

0-tyypin kieliopit ovat kielten määrittelymekanismeja, jotka ovat toimintatavaltaan pykälässä 1.3 määritellyistä tunnistusmekanismeista poiketen kielten generointimekanismeja. Kuten äärellisillä automaateilla, 0-tyypin kielioppien toiminta määräytyy äärellisiin komponentteihin sisältyvän informaation perusteella, mutta ne kykenevät prosessoimaan sanoja, mistä johtuen niiden laskentavoima on ratkaisevasti suurempi. Pykälässä 1.5 esitettävistä syistä 0-tyypin kieliopit edustavat korkeinta algoritmisen laskettavuuden tasoa.

0-tyypin kieliopin $G = (N, T, S, P)$ komponenteista N on aakkosto, jota kutsutaan *väliaakkostoksi* ja T on väliaakkostosta erillinen aakkosto, jota kutsutaan *pääteaakkostoksi*. Väli- ja pääteaakkostojen kirjaimia sanotaan *väli-* ja *päätesymboleiksi*. Komponentti $S \in N$ on väliaakkoston kirjain, jota kutsutaan *alkusymboliksi* ja

$$P \subseteq (N \cup T)^* N (N \cup T)^* \to (N \cup T)^*, \tag{20}$$

missä $\to$ on aakkostojen N ja T ulkopuolinen symboli, on äärellinen *johtosääntöjen* joukko. Kuten sisältymisen 20 oikeasta puolesta nähdään, 0-tyypin kieliopien johtosäännöissä $u \to v$ sana u sisältää aina vähintään yhden välisymbolin, mutta muuta rajoitusta johtosääntöjen muodolle ei ole. Joukon $(N \cup T)^*$ ehdosta

$$\begin{aligned} x \Rightarrow_G y \quad \Leftrightarrow \quad & x = x_1 u x_2 \in (N \cup T)^* \\ & y = x_1 v x_2 \in (N \cup T)^* \\ & u \to v \in P \end{aligned}$$

määräytyvä relaatio $\Rightarrow_G$ kuvaa kieliopin G yhtä toiminta-askeletta, jossa sanasta x *johdetaan* sana y.

Huomautus 1.14. Jos on selvää, mistä kieliopista on kyse, käytetään merkinnän $\Rightarrow_G$ sijasta merkintää $\Rightarrow$.

Relaation $\Rightarrow$ refleksiivisen ja transitiivisen sulkeuman $\Rightarrow^*$ avulla voidaan määritellä kieliopin G *generoima* kieli

$$L(G) = \{x \in T^* \mid S \Rightarrow^* x\}.$$

Se koostuu niistä sanoista yli pääteaakkoston, jotka voidaan johtaa soveltamalla alkusymboliin ja myöhemmin johdon tuloksena saatuihin sanoihin epädeterministisesti kieliopin G johtosääntöjä.

Kieliopin G *johdolla* tarkoitetaan jonoa

$$S \Rightarrow w_1, w_1 \Rightarrow w_2, w_2 \Rightarrow w_3, \ldots, w_{k-1} \Rightarrow w_k,$$

joka voidaan kirjoittaa myös muodossa

$$S \Rightarrow w_1 \Rightarrow w_2 \Rightarrow w_3 \Rightarrow \ldots \Rightarrow w_k,$$

missä lukua k sanotaan johdon *pituudeksi*.

Edellä esitetyn mukaan kielioppi G aloittaa toimintansa alkusymbolistaan S. Jos $w \in (N \cup T)^*$ on sana, joka toteuttaa ehdon $S \leq w$, kielioppi G voi aloittaa toimintansa sanasta w. Niiden sanojen joukosta yli pääteaakkoston T, jotka voidaan johtaa kieliopilla G lähtien sanasta w, käytetään merkintää

$$L(G, w) = \{x \in T^* \mid w \Rightarrow^* x\}.$$

Tällöin $L(G, S) = L(G)$.

Kieliopit G_1 ja G_2 ovat *ekvivalentit*, jos ne generoivat tarkalleen saman kielen eli

$$L(G_1) = L(G_2).$$

Niiden kielten luokasta, jotka voidaan generoida 0-tyypin kieliopeilla, käytetään merkintää $\mathcal{L}_0$.

Seuraavan Kuroda-normaalimuotona tunnetun lauseen todistuksessa on käytetty lähteenä kirjoja [10] ja [15].

Lause 1.15. *Jokaista 0-tyypin kielioppia G kohti on olemassa kieliopin G kanssa ekvivalentti kielioppi G', jolla on vain muotoa*

1. $A \to a, \quad a \in N, a \in T$
2. $A \to BC, \quad A, B, C \in N$
3. $AB \to CD, \quad A, B, C, D \in N$
4. $A \to \lambda, \quad A \in N$

olevia johtosääntöjä.

Todistus. Olkoon $G = (N, T, S, P)$ 0-tyypin kielioppi. Muodostetaan kielioppi $G_1 = (N_1, T, S, P_1)$, jolla on väliaakkosto

$$N_1 = N \cup \{A_a \mid a \in T\},$$

missä jokaista kirjainta $a \in T$ kohti otetaan käyttöön uusi välisymboli A_a ja jonka johtosäännöt P_1 saadaan johtosäännöistä P korvaamalla joukon P johtosäännöissä esiintyvät kirjaimet $a \in T$ kirjaimilla A_a sekä lisäämällä joukkoon P_1 säännöt

$$A_a \to a, \quad a \in T. \tag{21}$$

Näytetään, että kieliopit G_1 ja G ovat ekvivalentteja. Jos $\lambda \in L(G)$, niin korvaamalla kieliopin G johdossa

$$S \Rightarrow_G \ldots \Rightarrow_G \lambda \tag{22}$$

kirjaimet $a \in T$ kirjaimilla A_a saadaan kieliopin G_1 johto

$$S \Rightarrow_{G_1} \ldots \Rightarrow_{G_1} \lambda, \tag{23}$$

joten $\lambda \in L(G_1)$. Jos kääntäen $\lambda \in L(G_1)$, niin korvaamalla muotoa 23 olevassa johdossa kirjaimet A_a, $a \in T$ kirjaimilla a saadaan muotoa 22 oleva johto, joten $\lambda \in L(G)$. Jos $a_1 \cdots a_n \in L(G)$ on tyhjästä eroava sana, niin kieliopilla G_1 voidaan vastaavasti kuin edellä johtaa sana $A_{a_1} \cdots A_{a_n}$ ja soveltamalla siihen sääntöjä 21 sana $a_1 \cdots a_n$. Jos kääntäen $a_1 \cdots a_n \in L(G_1)$ on tyhjästä eroava sana, niin sana $a_1 \cdots a_n$ on saatu käyttämällä sääntöjä

$$P_1 \backslash \{A_a \to a \mid a \in T\} \tag{24}$$

ja muuttamalla jossain vaiheessa välisymbolit sääntöjen 21 avulla päätesymboleiksi. Sääntöjä 24 vastaavia sääntöjä käyttäen kieliopilla G voidaan johtaa sana $a_1 \cdots a_n$. Näin ollen $L(G_1) = L(G)$.

Muodostetaan kielioppi $G_2 = (N_1 \cup \{Y\}, T, S, P_2)$, missä Y on uusi välisymboli ja

$$\begin{aligned} P_2 \quad = \quad & \{u \to v \in P_1 \mid |u| \le |v|\} \cup \\ & \{u \to Y^{|u|-|v|}v \mid u \to v \in P_1,\ |u| > |v|\} \cup \{Y \to \lambda\}. \end{aligned}$$

Kieliopilla G_2 voidaan johtaa kaikki kielen $L(G_1)$ sanat käyttämällä sääntöjä $P_2 \backslash \{Y \to \lambda\}$ ja poistamalla mahdolliset kirjaimet Y säännöllä $Y \to \lambda$. Koska kirjain Y esiintyy vain säännön $Y \to \lambda$ vasemmalla puolella, kaikki kieliopilla G_2 johdetut joukkoon T^* kuuluvat sanat voidaan johtaa kieliopilla G_1, joten

$$L(G_2) = L(G_1). \tag{25}$$

Lisäksi sääntöä $Y \to \lambda$ lukuunottamatta kaikki kieliopin G_2 johtosäännöt ovat muotoa

$$u \to v, \quad |u| \le |v|.$$

Korvaamalla kieliopin G_2 muotoa

$$A \to B, \quad A, B \in N_1 \cup \{Y\}$$

olevat johtosäännöt säännöillä

$$A \to YB, \tag{26}$$

muotoa

$$A \to B_1 \cdots B_n, \quad n \ge 3$$

olevat säännöt säännöillä

$$\begin{array}{ccc} A & \to & B_1 Z_1 \\ Z_1 & \to & B_2 Z_2 \\ Z_2 & \to & B_3 Z_3 \\ & \vdots & \\ Z_{n-3} & \to & B_{n-2} Z_{n-2} \\ Z_{n-2} & \to & B_{n-1} B_n, \end{array} \tag{27}$$

missä kirjaimet $Z_1, \ldots, Z_{n-2}$ ovat uusia välisymboleja ja muotoa

$$A_1 \cdots A_n \rightarrow B_1 \cdots B_m, \quad 2 \leq n \leq m, (n,m) \neq (2,2)$$

olevat säännöt säännöillä

$$\begin{array}{ccc} A_1 A_2 & \rightarrow & B_1 Z_1' \\ Z_1' A_3 & \rightarrow & B_2 Z_2' \\ Z_2' A_4 & \rightarrow & B_3 Z_3' \\ & \vdots & \\ Z_{n-2}' A_n & \rightarrow & B_{n-1} Z_{n-1}' \\ Z_{n-1}' & \rightarrow & B_n Z_n' \\ Z_n' & \rightarrow & B_{n+1} Z_{n+1}' \\ & \vdots & \\ Z_{m-3}' & \rightarrow & B_{m-2} Z_{m-2}' \\ Z_{m-2}' & \rightarrow & B_{m-1} B_m, \end{array} \tag{28}$$

missä kirjaimet $Z_1', \ldots, Z_{m-2}'$ ovat uusia välisymboleja, ja lisäämällä kaikki prosessin aikana käyttöön otetut ja keskenään erilliset välisymbolit Z_i, Z_j' kieliopin G_2 välisymboleihin saadaan kielioppi G_3, jonka johtosäännöt ovat haluttua muotoa.

Vastaavalla päättelyllä kuin yhtälöä 25 perusteltaessa voidaan todeta, että säännöt 26 eivät muuta generoitua kieltä. Sääntöjä 27 käyttäen voidaan suorittaa johdot

$$A \Rightarrow_{G_3}^{*} B_1 \cdots B_n \tag{29}$$

ja välisymbolien Z_i erillisyydestä ja sääntöjen 27 muodosta seuraa, ettei sääntöjä 27 käyttäen voida johtaa kieleen $L(G_2)$ kuulumattomia sanoja. Sääntöjen 28 avulla voidaan suorittaa johdot

$$A_1 \cdots A_n \Rightarrow_{G_3}^{*} B_1 \cdots B_m, \; 2 \leq n \leq m, (n,m) \neq (2,2),$$

ja välisymbolien Z_i' erillisyydestä ja sääntöjen 28 muodosta seuraa, ettei sääntöjen 28 avulla voida johtaa kieleen $L(G_2)$ kuulumattomia sanoja. Näin ollen $L(G_3) = L(G_2) = L(G_1) = L(G)$. □

1.5 Laskettavuus ja universaalisuus

Sellaiset laskennalliset ongelmat, joissa kuhunkin ongelman tapaukseen tulee liittää joko myönteinen tai kielteinen vastaus, voidaan tulkita formaalisiksi kieliksi.

Esimerkki 1.16. Ongelma, jossa kysytään, onko annettu kokonaisluku neliö, voidaan tulkita kieleksi

$$L = \{a^{n^2} \mid n \in \mathbb{N}\}$$

yli aakkoston $\{a\}$. Ongelman tapauksia 0, 2, 5 ja 9 vastaavat kielen L sanat λ, a^2, a^5 ja a^9. Koska luvut 0 ja 9 ovat neliöitä, sanat λ ja a^9 kuuluvat kieleen L ja koska luvut 2 ja 5 eivät ole neliöitä, sanat a^2 ja a^5 eivät kuulu kieleen L.

Algoritmilla tarkoitetaan tehokasta menettelyä tai mekanismia jonkin laskennallisen ongelman ratkaisemiseksi. Sen jokainen askel suoritetaan äärellisen informaation perusteella, mutta käytössä olevaa aikaa ja talletustilaa ei rajoiteta. Jos rajoitutaan pykälän alussa mainittua tyyppiä oleviin ongelmiin, niin algoritmi voidaan formaalisten kielten teoriassa tulkita tehokkaaksi menettelyksi, joka ratkaistessaan ongelman luettelee sitä vastaavan kielen sanat.

Huomautus 1.17. Sanoja menettely ja mekanismi käytetään jatkossa synonyymeinä.

Huomautus 1.18. Lukua 3 lukuunottamatta, jossa sanalla mekanismi viitataan laskentamallien luokkaan, mekanismin komponenttien oletetaan olevan kiinteät.

Kieli $L \subseteq \Sigma^*$ on *rekursiivisesti numeroituva*, jos on olemassa tehokas menettely, joka luettelee tarkalleen kielen L sanat. Sen ei edellytetä pysähtyvän äärellisessä ajassa, koska kysymykseen voivat tulla äärettömät kielet. Rekursiivisesti numeroituvien kielten luokasta käytetään merkintää RE.

Huomautus 1.19. Edellinen rekursiivisesti numeroituvan kielen määritelmä ei ole matemaattisessa mielessä täsmällinen, koska tehokasta menettelyä ei voida määritellä matemaattisesti täsmällisesti.

Lause 1.20. *Luokan $\mathcal{L}_0$ kielet ovat rekursiivisesti numeroituvia.*

Todistus. Olkoon G 0-tyypin kielioppi. Seuraava menettely on tehokas menettely, joka luettelee tarkalleen kielen $L(G)$ sanat. Muodostetaan kaikki kieliopin G johdot pituuden mukaan järjestettynä siten, että ensin muodostetaan ne, joiden pituus on 1, sitten ne, joiden pituus on 2 jne. Samanpituiset johdot muodostetaan jossakin järjestyksessä, joka voidaan tehokkaasti toteuttaa. Jokaisen johdon yhteydessä tarkistetaan, syntyykö sen tuloksena sana yli pääteaakkoston. Myönteisessä tapauksessa kyseinen sana luetellaan kieleen $L(G)$ kuuluvaksi, muutoin siirrytään seuraavaan johtoon. $\square$

Määritellään käyttäen lähteenä kirjaa [10] tunnistusmekanismi, joka on äärellisten automaattien yleistys ja joka edustaa korkeinta algoritmisen laskettavuuden tasoa kaikkien laskennallisten ongelmien suhteen. *Turingin koneen*

$$M = (Q, \Sigma, \Gamma, \delta, \{q_0\}, B, F),$$

komponenteista $Q, \Sigma, \{q_0\}$ ja F ovat kuten äärellisillä automaateilla, Γ on *nauha-aakkosto*, joka sisältää aidosti syöteaakkoston Σ, $B \in \Gamma$ on aakkoston Σ ulkopuolinen kirjain ja

$$\delta : Q \times \Gamma \to \mathcal{P}(Q \times \Gamma \times \{L, R\})$$

on siirtymäfunktio, joka sallii epädeterministisyyden. Turingin koneella on käytössään yksi ruutuihin jaettu vasemmalta rajoitettu ja oikealle äärettömiin jatkuva *nauha* sekä yhden yhden kirjaimen lukemiseen ja kirjoittamiseen pystyvä *lukupää*. Turingin koneen syöte kirjoitetaan nauhan vasempaan päätyyn ja lukupää asetetaan syötteen vasemmanpuoleisimman kirjaimen kohdalle. Tällöin nauhan muut ruudut sisältävät kirjaimen B. Yhdessä toiminta-askelessaan Turingin kone lukee yhden kirjaimen lukupäänsä kohdalta, kirjoittaa jonkin kirjaimen kyseiseen ruutuun, siirtää lukupäätään yhden ruudun verran vasemmalle tai oikealle ja muuttaa tilaansa siirtymäfunktion δ ilmoittamalla tavalla.

Nauhan ja Turingin koneen tila tietyllä hetkellä voidaan ilmoittaa sanalla

$$\alpha q X \beta, \quad \alpha, \beta \in \Gamma^*, q \in Q, X \in \Gamma,$$

missä α on lukupään vasemmalla puolella olevista kirjaimista koostuva sana, $q \in Q$ on Turingin koneen tila, $X \in \Gamma$ on lukupään kohdalla olevan kirjain ja $\beta \in \Gamma^*$ on lukupään oikealla puolella olevista kirjaimista koostuva sana luettuna siihen saakka, kunnes nauhan loppuosassa on vain kirjaimia B. Joukon $\Gamma^* Q \Gamma^+$ ehdoista

$$\begin{aligned}
\alpha q X \beta \vdash_M \alpha Y p \beta \quad &\Leftrightarrow \quad (p, Y, R) \in \delta(q, X) \\
\alpha q \vdash_M \alpha Y p \quad &\Leftrightarrow \quad (p, Y, R) \in \delta(q, B) \\
\alpha Z q X \beta \vdash_M \alpha p Z Y \beta \quad &\Leftrightarrow \quad (p, Y, L) \in \delta(q, X) \\
\alpha Z q \vdash_M \alpha p Z Y \quad &\Leftrightarrow \quad (p, Y, L) \in \delta(q, B),
\end{aligned}$$

missä $\alpha, \beta \in \Gamma^*, X, Y, Z \in \Gamma, p, q \in Q$, määräytyvän relaation $\vdash_M$ refleksiivisen ja transitiivisen sulkeuman $\vdash_M^*$ avulla voidaan määritellä Turingin koneen M *tunnistama* kieli

$$L(M) = \{w \in \Sigma^* \mid q_0 w \vdash_M^* \alpha p \beta, p \in F, \alpha, \beta \in \Gamma^*\}.$$

Seuraavaa Churchin–Turingin teesin nimellä tunnettua väittämää ei voida todistaa, koska se koskee tehokkaita menettelyjä, joita ei voida määritellä matemaattisesti täsmällisesti.

Propositio 1.21. *Jokaista luokan RE kieltä L kohti on olemassa Turingin kone, jonka tunnistama kieli on L.*

Huomautus 1.22. Propositiota 1.21 voisi kutsua Churchin–Turingin teesiksi rajoittuen ongelmiin, joissa jokaiseen ongelman tapaukseen tulee liittää joko myönteinen tai kielteinen vastaus. Yleisemmässä muodossaan Churchin–Turingin teesi sanoo, että kaikki, mikä voidaan laskea algoritmisesti, voidaan laskea Turingin koneilla.

Huomautus 1.23. Proposition 1.21 mukaan luokka RE sisältyy Turingin koneilla tunnistettavissa olevien kielten luokkaan. Käänteinen sisältyminen voidaan todistaa, kuten on tehty luentomonisteessa [9], konstruoimalla useampinauhaisten Turingin koneiden avulla jokaista Turingin konetta M mekanismi joka luettelee kielen $L(M)$ sanat.

Seuraava lause on todistettu kirjassa [10].

Lause 1.24. *Jokaista Turingin konetta M kohti on olemassa 0-tyypin kielioppi G, jolle pätee $L(G) = L(M)$.*

Lauseen 1.24 ja proposition 1.21 nojalla 0-tyypin kielioppien voidaan katsoa edustavan korkeinta algoritmisen laskettavuuden tasoa, kun rajoitutaan pykälän alussa mainittua tyyppiä oleviin ongelmiin:

Lause 1.25. *Jokaista luokan RE kieltä L kohti on olemassa 0-tyypin kielioppi, joka toteuttaa $L(G) = L$.*

Olkoot m_1 ja m_2 mekanismeja, jotka määrittelevät kielet $L(m_1)$ ja $L(m_2)$. Mekanismit m_1 ja m_2 ovat *ekvivalentteja*, jos

$$L(m_1) = L(m_2).$$

Jos M on mekanismien luokka, niin luokan M mekanismi m_U on *universaali* luokan M suhteen, jos mekanismilla m_U voidaan simuloida mitä tahansa luokan M mekanismia m' siten, että mekanismi m_U määrittelee kielen $L(m')$, kun mekanismin m' toimintaa koskeva informaatio liitetään mekanismiin m_U.

Seuraavan lauseen mukaan pääteaakkostolla T varustettujen 0-tyypin kielioppien luokassa on universaali mekanismi. Lauseen todistus konstruktioineen löytyy kirjasta [12].

Lause 1.26. *On olemassa sellainen 0-tyypin kielioppi* $G_U = (N_U, T, S, P)$, *että jokaista pääteaakkostolla* T *varustettua 0-tyypin kielioppia* G *kohti on olemassa sellainen sana* $w(G) \in (N_U \cup T)^*$, *että* $L(G_U, w(G)) = L(G)$.

Universaali kielioppi G_U aloittaa toimintansa sanasta $w(G)$, joka sisältää kieliopin G_U väli- ja pääteaakkostojen yli koodattuna kieliopin G johtosäännöt. Kieliopilla G_U on johtosäännöt, joilla voidaan simuloida sanan $w(G)$ sisällä kieliopin G johtoja ja joilla voidaan poistaa sanasta $w(G)$ välisymbolit tarkalleen silloin, kun simuloinnin tuloksena syntyy sana yli pääteaakkoston T.

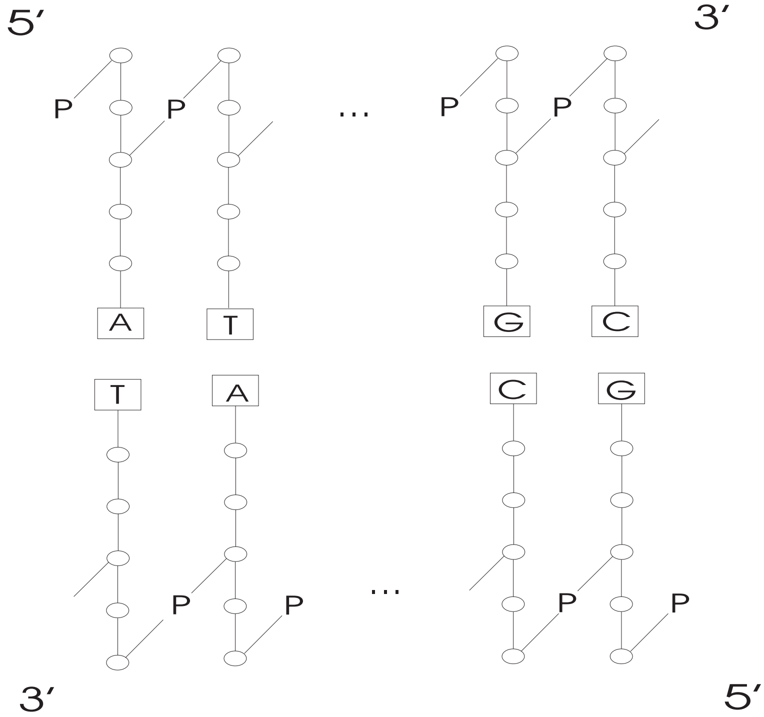
2 DNA formaalisten kielten teorian näkökulmasta

DNA:n kyky *kantaa* geneettistä informaatiota perustuu siihen, että sen neljä nukleotidiä voivat muodostaa erilaisia järjestyksiä. Koska nykyisin laboratoriotekniikoin voidaan konstruoida halutunlaisia nukleotidijonoja, DNA-molekyylejä voidaan käyttää informaation tallennusmuotona. Tämä on lähtökohtana DNA-laskennalle - ratkaistava ongelma tulee ensin koodata DNA-molekyyliksi tai DNA-molekyylien joukoksi. Luvun alussa annetaan DNA-molekyyleille esitys sanoina, jolloin DNA-molekyylien joukkoja voidaan tarkestella kielinä.

DNA:n kyvyn *kopioida* geneettistä informaatiota mahdollistaa nukleotidien pariutumisen määräävä komplementaarisuus. Luvun lopussa tarkastellaan komplementaarisuuden laskentavoimaa jonokoneisiin yhdistettynä, joka osoittautuu ekvivalentiksi Turingin koneiden kanssa.

2.1 DNA-molekyylien esitys sanoina

Tarkastellaan aluksi DNA-molekyylien rakennetta rajoittuen *lineaarisiin* molekyyleihin eli molekyyleihin, jotka eivät muodosta ympyrää eivätkä haaraudu. *DNA-kaksoiskierre* koostuu kahdesta toistensa ympäri kiertyneestä *säikeestä*, jotka puolestaan koostuvat toisiinsa kiinnittyneistä *nukleotideista*. Nukleotidi on viiden hiiliatomin ketju, jonka toiseen päähän on liittynyt emäs ja toiseen fosfaatti. Nukleotidejä on neljä erilaista ja ne eroavat toisistaan ainoastaan emäksen osalta. Nukleotideille on annettu nimet adeniini, guaniini, sytosiini ja tymiini, ja on tapana merkitä niitä kirjaimilla A,G,C ja T. Nukleotidi voi liittyä toiseen joko emästen ollessa kohdakkain, jolloin syntyvä sidos on heikko, tai vahvalla fosfaatin ja emäksestä laskettuna kolmannen hiilen välisellä sidoksella. Emästen kautta syntyvä sidos on eri säikeisiin kuuluvien nukleotidien välillä ja fosfaatin ja hiilen välinen sidos on samaan säikeeseen kuuluvien nukleotidien välillä. Emästen välisen sidoksen heikkoudesta johtuen kaksoiskierre voidaan erottaa kahdeksi säikeeksi kuumentamalla, kun taas samaan säikeeseen kuuluvien nukleotidien välisen sidoksen vahvuudesta johtuen säie on vankkatekoisempi. Kuva 1, jossa nukleotidien fosfaattia on merkitty kirjaimella P, havainnollistaa DNA-kaksoiskierteen rakennetta.



Kuva 1: DNA-kaksoiskierteen rakenne

Huomautus 2.1. Jatkossa nukleotidin kiinnittymisellä toiseen nukleotidiin tarkoitetaan säikeiden välistä sidosta. Nukleotidin x viereisellä nukleotidilla tarkoitetaan nukleotidin x kanssa samaan säikeeseen kuuluvaa nukleotidiä, jolla on yhteinen sidos nukleotidin x kanssa.

Säikeen laitimmaiset nukleotidit eroavat toisistaan siten, että vain toisella on fosfaatti, joka ei osallistu sidokseen. On tapana käyttää säikeen vapaalla fosfaatilla varustetusta päästä merkintää 5' ja sidokseen osallistuvalla fosfaatilla varustetusta päästä merkintää 3'. Katenoimalla nukleotidejä vastaavat kirjaimet alkaen 5'-päästä 3'-päähän saakka saadaan tyhjästä eroava sana w yli aakkoston

$$\Sigma_{DNA} = \{A, G, C, T\},$$

jonka voidaan tulkita esittävän kyseistä säiettä. Kääntäen jokaisen sanan $w \in \Sigma_{DNA}^{+}$ voidaan tulkita esittävän säiettä, jonka nukleotidit on lueteltu alkaen 5'-päästä. Tyhjällä sanalla ei ole esitystä säikeenä, mutta koska tyhjä sana kuuluu morfismien määrittelyjoukkoon, se voidaan ottaa tarkasteluihin mukaan.

Kaksoiskierteessä adeniini on aina kiinnittynyt tymiiniin ja guaniini sytosiiniin kuten kuvassa 1. Tätä DNA:n ominaisuutta sanotaan *komplementaarisuudeksi*. Määritellään morfismi $c : \Sigma_{DNA}^{*} \to \Sigma_{DNA}^{*}$,

$$\begin{array}{ll} c(A) = T & c(G) = C \\ c(T) = A & c(C) = G, \end{array} \tag{30}$$

joka kuvaa jokaisen nukleotidin $x \in \Sigma_{DNA}$ siksi nukleotidiksi, johon nukleotidi x on kiinnittynyt kaksoiskierteessä. Koska

$$c(c(x)) = x \quad \forall x \in \Sigma_{DNA},$$

morfismille c pätee

$$c(c(w)) = w \quad \forall w \in \Sigma_{DNA}^{*}. \tag{31}$$

Sanasta $c(w)$ käytetään jatkossa merkintää $\overline{w}$, ja sitä käyttäen yhtälö 31 saa muodon

$$\overline{\overline{w}} = w \quad \forall w \in \Sigma_{DNA}^{*}. \tag{32}$$

Kuten kuvassa 1, kaksoiskierteen molemmissa päissä 5'-nukleotidi on aina kiinnittynyt 3'-nukleotidiin eli säikeet kulkevat eri suuntiin suuntien $5' \to 3'$ ja $3' \to 5'$ suhteen. Näin ollen sana $\overline{w^R}$ tarkoittaa säikeeseen $w \in \Sigma_{DNA}^{+}$ kiinnittynyttä säiettä. Jos $w = a_1 a_2 \cdots a_k$, missä $a_i \in \Sigma_{DNA}$, kun $1 \leq i \leq k$, niin

$$\begin{aligned} \overline{w^R} &= \overline{(a_1 a_2 \ldots a_k)^R} \\ &= \overline{(a_k a_{k-1} \ldots a_1)} \\ &= \overline{a_k}\,\overline{a_{k-1}} \ldots \overline{a_1} \\ &= (\overline{a_1}\,\overline{a_2} \ldots \overline{a_k})^R \\ &= (\overline{a_1\, a_2 \ldots a_k})^R \\ &= \overline{w}^R, \end{aligned} \tag{33}$$

joten säikeeseen w kiinnittyneestä säikeestä voidaan käyttää myös merkintää $\overline{w}^R$. Kuten pitääkin, siirtymällä säikeeseen w kiinnittyneestä säikeestä $\overline{w^R}$ säikeeseen $\overline{w^R}$ kiinnittyneeseen säikeeseen $\overline{(\overline{w^R})^R}$, päädytään yhtälöjen 33, 32 ja peilikuvaoperaation ominaisuuden 1 nojalla takaisin säikeeseen w:

$$\begin{aligned}\overline{(\overline{w^R})^R} &= (\overline{\overline{w^R}})^R \\ &= (w^R)^R \\ &= w.\end{aligned}$$

Komplementaarisuuden takia kaksoiskierre määräytyy täysin yhden säikeen perusteella. Jotta saataisiin tehtyä merkinnällinen ero säikeiden ja kaksoiskierteiden välille, esitetään jatkossa säikeitä sanoilla

$$w \in \Sigma_{DNA}^+$$

ja kaksoiskierteitä muotoa

$$(w, \overline{w^R}), \quad w \in \Sigma_{DNA}^+$$

olevilla pareilla, missä sana w on jompaakumpaa kaksoiskierteen säikeistä vastaava sana. Jos tarkasteltaisiin vain kaksoiskierteiden joukkoja eikä sekaannusvaaraa olisi, voitaisiin kaksoiskierteitä esittää sanoina yli aakkoston Σ_{DNA}. Tällöin jokaista kaksoiskierrettä vastaisivat sanat w ja $\overline{w^R}$.

Huomautus 2.2. Komplementaarisuudesta johtuen kaikilla kaksoiskierteillä $(w, \overline{w^R})$ kaksoiskierteen sisältämien nukleotidien A, G ja C, T lukumäärien suhteelle pätee

$$\frac{|w|_A + |\overline{w^R}|_A + |w|_G + |\overline{w^R}|_G}{|w|_C + |\overline{w^R}|_C + |w|_T + |\overline{w^R}|_T} = \frac{|w|_A + |w|_T + |w|_G + |w|_C}{|w|_C + |w|_G + |w|_T + |w|_A} = 1.$$

Sen sijaan kirjan [19] mukaan kaksoiskierteen sisältämien nukleotidien A, T ja G, C lukumäärien suhde

$$\begin{aligned}\frac{|w|_A + |\overline{w^R}|_A + |w|_T + |\overline{w^R}|_T}{|w|_G + |\overline{w^R}|_G + |w|_C + |\overline{w^R}|_C} &= \frac{|w|_A + |w|_T + |w|_T + |w|_A}{|w|_G + |w|_C + |w|_C + |w|_G} \\ &= \frac{|w|_A + |w|_T}{|w|_G + |w|_C}\end{aligned}$$

vaihtelee, mutta on suurinpiirtein vakio samaan lajiin kuuluvien eliöiden soluista tavattavilla kaksoiskierteillä.

2.2 Biologiset palindromit

Muotoa (w, w), missä $w \in \Sigma_{DNA}^+$, olevat kaksoiskierteet toteuttavat ehdon

$$w = \overline{w^R}. \tag{34}$$

Kyseistä muotoa olevat kaksoiskierteet näyttävät samalta sekä ylä- että alapuolelta katsottaessa. Ehdon 34 täyttäviä sanoja $w \in \Sigma_{DNA}^*$, joihin tyhjä sana selvästi lukeutuu, sanotaan *biologisiksi palindromeiksi*. Koska morfismi 30 ei pidä mitään aakkoston Σ_{DNA} kirjainta paikallaan, mikään tyhjästä sanasta eroava biologinen palindromi ei ole palindromi.

Huomautus 2.3. Jos morfismi c korvattaisiin identiteettikuvauksella, biologiset palindromit olisivat tarkalleen palindromit. Morfismin c korvaaminen identiteettikuvauksella olisi perusteltua, jos komplementaarisuudelta vaadittaisiin vain yhtälön 31 toteutumista, koska selvästi

$$Id \circ Id = Id.$$

Huomautus 2.4. Artikkelin [18] mukaan luonnossa esiintyvien DNA-molekyylien proteiineja koodaavat alueet sisältävät biologisia palindromeja.

Seuraavan lemman nojalla biologiset palindromit ovat parillispituisia.

Lemma 2.5. *Sana $w \in \Sigma_{DNA}^*$ on biologinen palindromi tarkalleen silloin, kun $w = u\overline{u}^R$ jollakin sanalla $u \in \Sigma_{DNA}^*$.*

Todistus. Jos $w = u\overline{u}^R$, niin peilikuvaoperaation ominaisuuksien 2 ja 1 ja yhtälöiden 32 ja 33 nojalla

$$\begin{aligned}
\overline{w^R} &= \overline{(u\overline{u}^R)^R} \\
&= \overline{(\overline{u}^R)^R u^R} \\
&= \overline{\overline{u} u^R} \\
&= \overline{\overline{u}}\,\overline{u^R} \\
&= u\overline{u}^R \\
&= w.
\end{aligned}$$

Jos kääntäen sana $w \in \Sigma_{DNA}^*$ on biologinen palindromi eli

$$w = \overline{w^R}, \tag{35}$$

niin sanan w täytyy olla parillispituinen, sillä jos $|w|$ olisi pariton, niin sanan w keskimmäisellä kirjaimella a pätisi $a = \overline{a}$, mikä ei ole mahdollista morfismin 30 määrittelyn nojalla. Sana w voidaan siis kirjoittaa muotoon

$$w = uv, \quad |u| = |v|. \tag{36}$$

Yhtälöiden 36, 35, peilikuvaoperaation ominaisuuden 1 ja yhtälön 33 nojalla

$$\begin{aligned} uv &= w \\ &= \overline{w^R} \\ &= \overline{(uv)^R} \\ &= \overline{v^R u^R} \\ &= \overline{v^R}\,\overline{u^R} \\ &= \overline{v}^R \overline{u}^R. \end{aligned}$$

Yhtälöistä $|u| = |v|$ ja $uv = \overline{v}^R \overline{u}^R$ voidaan päätellä, että $v = \overline{u}^R$, jolloin

$$\begin{aligned} w &= uv \\ &= u\overline{u}^R. \end{aligned}$$

□

Esimerkki 2.6. Kieliopin

$$G = (\{S\}, \{A, G, T, C\}, S, \{S \to AST, S \to TSA, S \to GSC, S \to CSG, S \to \lambda\})$$

generoima kieli on biologiset palindromit. Johdon

$$S \Rightarrow AST \Rightarrow ATSAT \Rightarrow ATCSGAT \Rightarrow ATCGSCGAT \Rightarrow ATCGCGAT$$

tuloksena saadaan sana

$$ATCGCGAT = ATCG(TAGC)^R = ATCG(\overline{ATCG})^R.$$

2.3 DNA-kaksoiskierteiden yhteys TS- ja RTS-kieliin

Pykälässä 2.1 formalisoitiin DNA:n komplementaarisuus morfismin 30 avulla. Aakkosto Σ_{DNA} voidaan partitioida kahteen osaan $\{A, G\}$ ja $\{C, T\}$ siten, että jokainen aakkoston $\{A, G\}$ kirjain on tarkalleen yhden aakkoston $\{C, T\}$ kirjaimen komplementti ja kääntäen:

$$\begin{aligned} A = \overline{T}, &\quad C = \overline{G} \\ G = \overline{C}, &\quad T = \overline{A}. \end{aligned}$$

Jos nimetään aakkoston Σ_{DNA} kirjaimet uudelleen säännön

$$\begin{aligned} A \mapsto 0 &\quad G \mapsto 1 \\ T \mapsto \overline{0} &\quad C \mapsto \overline{1} \end{aligned}$$

mukaan, aakkoston Σ_{DNA} partitiota $\{A, G\}, \{T, C\}$ vastaa aakkoston $\{0, 1, \overline{0}, \overline{1}\}$ partitio $\{0, 1\}, \{\overline{0}, \overline{1}\}$.

Määritellään yleinen komplementaarisuusmorfismilla varustettu aakkosto. Olkoon $\Sigma \cup \overline{\Sigma}$ erillisten aakkostojen Σ ja $\overline{\Sigma} = \{\overline{a} \mid a \in \Sigma\}$ unioni ja $c : (\Sigma \cup \overline{\Sigma})^* \to (\Sigma \cup \overline{\Sigma})^*$ morfismi

$$\begin{aligned} c(a) &= \overline{a} \quad \forall a \in \Sigma \\ c(\overline{a}) &= a \quad \forall \overline{a} \in \overline{\Sigma}. \end{aligned}$$

Kun lisäksi merkitään

$$c(a) = \overline{a} \quad \forall a \in (\Sigma \cup \overline{\Sigma}),$$

jolloin $\overline{\overline{a}} = a$, morfismi c toteuttaa yhtälöjä 32 ja 33 vastaavat yhtälöt.

Määritellään kielet

$$\begin{aligned} TS_\Sigma &= \bigcup_{w \in \Sigma^*} w \sqcup\!\sqcup \overline{w} \\ RTS_\Sigma &= \bigcup_{w \in \Sigma^*} w \sqcup\!\sqcup \overline{w^R}, \end{aligned}$$

joille selvästi pätee

$$TS_\Sigma = \{w \in (\Sigma \cup \overline{\Sigma})^* \mid \overline{pr_\Sigma(w)} = pr_{\overline{\Sigma}}(w)\}, \tag{37}$$

$$RTS_\Sigma = \{w \in (\Sigma \cup \overline{\Sigma})^* \mid \overline{pr_\Sigma(w)} = (pr_{\overline{\Sigma}}(w))^R\}, \tag{38}$$

missä pr_Σ ja $pr_{\overline{\Sigma}}$ ovat projektiokuvaukset aakkostolta $\Sigma \cup \overline{\Sigma}$ aakkostoille Σ ja $\overline{\Sigma}$.

Seuraavilla algoritmeilla kaksoiskierteistä

$$(w, \overline{w^R}), \quad w \in \{0, 1, \overline{0}, \overline{1}\}^+$$

saadaan kielten $TS_{\{0,1\}}$ ja $RTS_{\{0,1\}}$ sanoja.

1. Luetellaan kaksoiskierteen

$$(w, \overline{w^R}) = (a_1 a_2 \cdots a_n, \overline{a_n}\,\overline{a_{n-1}} \cdots \overline{a_1}), \quad a_i \in \{0, 1, \overline{0}, \overline{1}\},\ 1 \le i \le n, \tag{39}$$

kirjaimet järjestyksessä

$$a_1, \overline{a_1}, a_2, \overline{a_2}, \ldots, a_n, \overline{a_n}$$

eli edetään säiettä w suuntaan $5' \to 3'$ ja säiettä $\overline{w^R}$ suuntaan $3' \to 5'$. Katenoimalla kirjaimet saadaan sana

$$a_1\overline{a_1}a_2\overline{a_2} \cdots a_n\overline{a_n} \in a_1 a_2 \cdots a_n \sqcup\!\sqcup \overline{a_1 a_2 \cdots a_n} \subseteq TS_{\{0,1\}},$$

joka kuuluu kieleen $TS_{\{0,1\}}$.

2. Luetellaan kaksoiskierteen 39 kirjaimet järjestyksessä

$$a_1, \overline{a_n}, a_2, \overline{a_{n-1}}, \ldots, a_{n-1}, \overline{a_2} a_n, \overline{a_1}$$

eli edeten molempia säikeitä suuntaan $5' \rightarrow 3'$. Katenoimalla kirjaimet saadaan sana

$$a_1 \overline{a_n} a_2 \overline{a_{n-1}} \cdots a_{n-1} \overline{a_2} a_n \overline{a_1} \in a_1 a_2 \cdots a_n \sqcup\!\sqcup \overline{(a_1 a_2 \cdots a_n)^R} \subseteq RTS_{\{0,1\}},$$

joka kuuluu kieleen $RTS_{\{0,1\}}$.

Koska algoritmilla 1 tuotettujen sanojen kaksi ensimmäistä kirjainta ovat aina toistensa komplementteja, ei ole olemassa kaksoiskierrettä, josta algoritmi 1 antaisi sanan

$$00\overline{00} \in 00 \sqcup\!\sqcup \overline{00} \subseteq TS_{\{0,1\}}.$$

Algoritmilla 2 tuotettujen sanojen muodosta nähdään, että ne ovat biologisia palindromeja. Koska sana $0\overline{00}0$ ei ole biologinen palindromi, ei ole olemassa kaksoiskierrettä, josta algoritmi 2 antaisi sanan

$$0\overline{00}0 \in 00 \sqcup\!\sqcup \overline{(00)^R} \subseteq RTS_{\{0,1\}}.$$

Esimerkki 2.7. Algoritmi 2 antaa kaksoiskierteestä

$$(w, \overline{w^R}) = (0\overline{01}11100\overline{1}1, \overline{1}1\overline{00111}10\overline{0})$$

biologisen palindromin

$$0\overline{10}1\overline{10}1\overline{0}1\overline{1}1\overline{1}0\overline{1}01\overline{1}01\overline{0},$$

joka eroaa sanasta

$$w\overline{w^R} = 0\overline{01}11100\overline{1}1\overline{1}1\overline{00111}10\overline{0},$$

joka lemman 2.5 nojalla on myös biologinen palindromi.

Seuraavilla algoritmeilla saadaan kaksoiskierteistä kaikki kielten $TS_{\{0,1\}}$ ja $RTS_{\{0,1\}}$ sanat. Nimetään ensin kaksoiskierteen $(w, \overline{w^R})$ kirjaimet uudelleen säännön

	Parin 1. säie	Parin 2. säie
A,T	0	$\overline{0}$
G,C	1	$\overline{1}$

mukaan.

3. Luetellaan ylläkuvatun säännön mukaan koodatun kaksoiskierteen

$$(w, \overline{w^R}), \quad w \in \{0,1\}^+, \overline{w^R} \in \{\overline{0}, \overline{1}\}$$

kirjaimet järjestyksessä

$$x_1, y_1, x_2, y_2, \ldots, x_n, y_n, \quad x_i \in \{0, 1, \lambda\}, y_i \in \{\overline{0}, \overline{1}, \lambda\}, i = 1, 2, \ldots, n \quad (40)$$

siten, että

$$\begin{aligned} x_1 x_2 \cdots x_n &= w \\ y_1 y_2 \cdots y_n &= \overline{w} \end{aligned}$$

eli edetään säiettä w suuntaan $5' \to 3'$ ja säiettä $\overline{w^R}$ suuntaan $3' \to 5'$. Katenoimalla kirjaimet saadaan sana

$$x_1 y_1 x_2 y_2 \cdots x_n y_n \in w \sqcup\!\sqcup\, \overline{w} \subseteq TS_{\{0,1\}}.$$

Käymällä epädeterministisesti läpi kaikki järjestykset 40 saadaan tuotettua kaikki joukon $w \sqcup\!\sqcup\, \overline{w}$ sanat. Koska jokaista sanaa $x \in \{0,1\}^+$ kohti voidaan konstruoida kaksoiskierre $(w, \overline{w^R})$, jossa parin ensimmäistä säiettä w vastaa sana x, voidaan algoritmilla 3 tuottaa kaikki kielen $TS_{\{0,1\}}$ sanat.

4. Muuttamalla algoritmissa 3 molempien säikeiden etenemissuunnaksi $5' \to 3'$ saadaan algoritmi, jolla voidaan kaksoiskierteistä tuottaa kaikki kielen $RTS_{\{0,1\}}$ sanat.

Esimerkki 2.8. Algoritmi 3 antaa kaksoiskierteestä (AT, AT), jota vastaa pari $(00, \overline{00})$, järjestyksellä

$$0, \lambda, 0, \lambda, \lambda, \overline{0}, \lambda, \overline{0}$$

sanan $00\overline{00}$.

Algoritmi 4 antaa samasta kaksoiskierteestä järjestyksellä

$$0, \overline{0}, \lambda, \overline{0}, 0, \lambda$$

sanan $0\overline{00}0$.

2.4 Komplementaarisuuden laskentavoima

Seuraavan lauseen todistuksessa on käytetty lähteenä kirjaa [16].

Lause 2.9. *Jokaista luokan RE kieltä yli aakkoston T kohti on olemassa sellaiset morfismit h_1, h_2 ja säännöllinen kieli R, että*

$$L = pr_T(E(h_1, h_2) \cap R).$$

Todistus. Olkoon $G = (N, T, S, P)$ mielivaltainen pääteaakkostolla T varustettu 0-tyypin kielioppi. Tulkitaan joukko P aakkostoksi, jolloin sanat

$$u \to v \in P \subseteq (N \cup T \cup \{\to\})^*$$

ovat kirjaimia. Määritellään aakkostot

$$\begin{aligned} T' &= \{a' \mid a \in T\} \\ \Sigma_1 &= N \cup T \cup \{\#\} \\ \Sigma_2 &= \Sigma_1 \cup P \cup T' \cup \{B, F\} \\ V_1 &= N \cup T', \end{aligned}$$

missä $\#$, B ja F ovat aakkostojen N, T, N', T', P ulkopuolisia kirjaimia, morfismit $h_1, h_2 : \Sigma_2^* \to \Sigma_1^*$,

	B	$\#$	$(\alpha_i \to \beta_i) \in P$	$A \in N$	$a' \in T'$	$a \in T$	F
h_1	$S\#$	$\#$	β_i	A	a	λ	λ
h_2	λ	$\#$	α_i	A	a	a	$\#$

ja kieli

$$R = B(V_1^* P V_1^* \#)^+ T^* F,$$

joka lauseen 1.6 ja lemman 1.7 nojalla on säännöllinen.

Muotoillaan apuväite, jonka mukaan jokaista kieliopin G johtoa

$$S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_k$$

kohti on olemassa sana $w \in B(V_1^* P V_1^* \#)^+$, joka toteuttaa ehdot

$$\begin{aligned} h_1(w) &= S\#\alpha_1\#\alpha_2\#\ldots\#\alpha_k\# && (41) \\ h_2(w) &= S\#\alpha_1\#\alpha_2\#\ldots\#\alpha_{k-1}\#. && (42) \end{aligned}$$

Todistetaan apuväite induktiolla johdon pituuden k suhteen. Jos $k = 1$, niin voidaan valita sana

$$w' = BS \to \alpha_1\#,$$

missä $S \to \alpha_1 \in P$. Tällöin nimittäin morfismien h_1 ja h_2 määrittelyn perusteella

$$\begin{aligned} h_1(w') &= h_1(B)h_1(S \to \alpha_1)h_1(\#) \\ &= S\#\alpha_1\# \\ h_2(w') &= h_2(B)h_2(S \to \alpha_1)h_2(\#) \\ &= S\#. \end{aligned}$$

Oletetaan, että väite pätee argumentin arvolla k ja johtoa

$$S \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_k \tag{43}$$

kohti on olemassa ehdot 41 ja 42 täyttävä sana w. Jatketaan johtoa 43 askeleella $\alpha_k \Rightarrow \alpha_{k+1}$, missä

$$\alpha_k = \beta_1 \alpha \beta_2 \tag{44}$$

$$\alpha_{k+1} = \beta_1 \beta \beta_2 \tag{45}$$

eli sovelletaan sanaan α_k sääntöä $\alpha \to \beta \in P$. Merkitään sanoilla β_1' ja β_2' sanoja, jotka saadaan sanoista β_1 ja β_2 lisäämällä pilkut joukon T päätesymboleihin. Sana

$$w\beta_1'\alpha \to \beta\beta_2'\#$$

on etsitty sana, sillä induktio-oletuksen, morfismien h_1, h_2 määrittelyn ja yhtälöjen 45 ja 44 nojalla

$$\begin{aligned}
h_1(w\beta_1'\alpha \to \beta\beta_2'\#) &= h_1(w)h_1(\beta_1')h_1(\alpha \to \beta)h_1(\beta_2')h_1(\#) \\
&= S\#\alpha_1\# \cdots \#\alpha_k\#h_1(\beta_1')h_1(\alpha \to \beta)h_1(\beta_2')h_1(\#) \\
&= S\#\alpha_1\# \ldots \#\alpha_k\#\beta_1\beta\beta_2\# \\
&= S\#\alpha_1\# \ldots \#\alpha_k\#\alpha_{k+1}\#, \\
h_2(w\beta_1'\alpha \to \beta\beta_2'\#) &= h_2(w)h_2(\beta_1')h_2(\alpha \to \beta)h_2(\beta_2')h_2(\#) \\
&= S\#\alpha_1\# \cdots \#\alpha_{k-1}\#h_2(\beta_1')h_2(\alpha \to \beta)h_2(\beta_2')h_2(\#) \\
&= S\#\alpha_1\# \ldots \#\alpha_{k-1}\#\beta_1\alpha\beta_2\# \\
&= S\#\alpha_1\# \ldots \#\alpha_{k-1}\#\alpha_k\#.
\end{aligned}$$

Näin ollen apuväite pätee kaikilla kieliopin G johdoilla.

Olkoon $x \in L$. Jos sana x on johdon

$$S \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_k,$$

missä $\alpha_k = x$, tulos, niin apuväitteen nojalla on olemassa sellainen sana $w \in B(V_1^*PV_1^*\#)^+$, että

$$\begin{aligned}
h_1(w) &= S\#\alpha_1\# \cdots \#\alpha_k\# \\
h_2(w) &= S\#\alpha_1\# \cdots \#\alpha_{k-1}\#.
\end{aligned} \tag{46}$$

Tällöin morfismien h_1 ja h_2 määrittelyn perusteella sanalle $wxF \in R$ pätee

$$\begin{aligned}
h_1(wxF) &= h_1(w)h_1(x)h_1(F) \\
&= S\#\alpha_1\# \cdots \#\alpha_k\#h_1(x)h_1(F)
\end{aligned}$$

$$
\begin{aligned}
&= S\#\alpha_1\#\cdots\#\alpha_k\# \\
h_2(wxF) &= h_2(w)h_2(x)h_2(F) \\
&= S\#\alpha_1\#\ldots\#\alpha_{k-1}\#h_2(x)h_2(F) \\
&= S\#\alpha_1\#\ldots\#\alpha_{k-1}\#x\# \\
&= S\#\alpha_1\#\ldots\#\alpha_{k-1}\#\alpha_k\#,
\end{aligned}
$$

joten $wxF \in E(h_1, h_2) \cap R$. Aakkostojen $\{B, \#\} \cup V_1 \cup P$ ja T erillisyyden nojalla

$$x = pr_T(wxF) \in pr_T(E(h_1, h_2) \cap R).$$

Näin ollen $L \subseteq pr_T(E(h_1, h_2) \cap R)$.

Näytetään, että $pr_T(E(h_1, h_2) \cap R) \subseteq L$. Olkoon $x = pr_T(w)$, missä $w \in E(h_1, h_2) \cap R$. Koska $w \in R$, sana w on muotoa

$$w = Bw_1w_2\ldots w_kx'F,\ w_i \in V_1^*PV_1^*\#,\ 1 \leq i \leq k, x' \in T^*.$$

Koska $x = pr_T(w)$ ja aakkostot $\{B, \#\} \cup V_1 \cup P$ ja T ovat erillisiä, $x' = x$ eli sana x sijaitsee sanassa w ennen kirjainta F. Kuvausten h_1 ja h_2 määrittelyn perusteella

$$
\begin{aligned}
h_1(Bw_1w_2\cdots w_k) &= S\#w_1^{(1)}\#w_2^{(1)}\#\cdots\#w_k^{(1)}\#, \\
h_2(Bw_1w_2\cdots w_k) &= w_1^{(2)}\#w_2^{(2)}\#\ldots\#w_k^{(2)}\#,
\end{aligned}
$$

missä sanat

$$w_i^{(1)},\ w_i^{(2)}, \quad i = 1, 2, \ldots, k$$

eivät sisällä kirjainta $\#$. Koska

$$h_1(Bw_1w_2\cdots w_kxF) = h_2(Bw_1w_2\cdots w_kxF), \tag{47}$$

jompikumpi relaatioista

$$
\begin{aligned}
h_1(Bw_1w_2\cdots w_k) &\leq h_2(Bw_1w_2\cdots w_k) \\
h_2(Bw_1w_2\cdots w_k) &\leq h_1(Bw_1w_2\cdots w_k)
\end{aligned}
$$

pätee. Asettamalla kirjaimet $\#$ kohdakkain nähdään, että

$$h_2(Bw_1w_2\cdots w_k) < h_1(Bw_1w_2\cdots w_k)$$

ja lisäksi saadaan yhtälöt

$$
\begin{aligned}
w_1^{(2)} &= S \\
w_i^{(2)} &= w_{i-1}^{(1)},\ i = 2, 3, \ldots, k.
\end{aligned}
$$

Jakamalla yhtälön 47 molemmilta puolilta yhteinen etuliite $S\#w_1^{(1)}\#\cdots\#w_{k-1}^{(1)}\#$ saadaan yhtälö

$$w_k^{(1)}\#h_1(xF) = h_2(xF), \tag{48}$$

ja sijoittamalla morfismien h_1 ja h_2 arvot yhtälöön 48 saadaan yhtälö

$$w_k^{(1)}\# = x\#,$$

joten $w_k^{(1)} = x$. Yhtälöistä

$$\begin{aligned} h_1(Bw_1\cdots w_k) &= S\#w_1^{(1)}\#w_2^{(1)}\#\cdots\#w_{k-1}^{(1)}\#w_k^{(1)}\#, \\ h_2(Bw_1\cdots w_k) &= S\#w_1^{(1)}\#w_2^{(1)}\#\cdots\#w_{k-1}^{(1)}\# \end{aligned}$$

seuraavista yhtälöistä

$$\begin{aligned} h_2(w_1) &= S\# \\ h_1(w_1) &= w_1^{(1)}\# \\ h_2(w_2) &= w_1^{(1)}\# \\ h_1(w_2) &= w_2^{(1)}\# \\ &\vdots \\ h_2(w_k) &= w_{k-1}^{(1)}\# \\ h_1(w_k) &= w_k^{(1)}\# \end{aligned} \tag{49}$$

seuraa morfismien h_1 ja h_2 määrittelyn perusteella, että on olemassa kieliopin G johto

$$S \Rightarrow w_1^{(1)} \Rightarrow w_2^{(1)} \Rightarrow \ldots \Rightarrow w_{k-1}^{(1)} \Rightarrow w_k^{(1)} = x.$$

Näin ollen $x \in L$. □

Seuraavan lauseen todistuksessa on käytetty lähteenä kirjaa [16].

Lause 2.10. *Jokaista luokan RE kieltä $L \subseteq T^*$ kohti on olemassa sellainen gsm-kuvaus $g : \{0, 1, \overline{0}, \overline{1}\} \to T^*$, että $L = g(TS_{\{0,1\}})$.*

Todistus. Olkoon $L \in RE$, $L \subseteq T^*$. Lauseen 2.9 mukaan on olemassa sellaiset morfismit $h_1, h_2 : {\Sigma_2}^* \to {\Sigma_1}^*$ ja säännöllinen kieli $R \subseteq \Sigma_2^*$, että

$$L = pr_T(E(h_1, h_2) \cap R).$$

Nimetään aakkoston Σ_1 kirjaimet uudelleen siten, että

$$\Sigma_1 \cap \Sigma_2 = \emptyset, \tag{50}$$

mikä ei vaikuta lauseen 2.9 todistukseen, ja määritellään aakkostot

$$\begin{aligned}\Sigma &= \Sigma_1 \cup \Sigma_2 \\ \overline{\Sigma} &= \{\overline{a} \mid a \in \Sigma\} \\ \overline{\Sigma_1} &= \{\overline{a} \mid a \in \Sigma_1\} \\ \overline{\Sigma_2} &= \{\overline{a} \mid a \in \Sigma_2\},\end{aligned}$$

morfismi $g : {\Sigma_2}^* \to (\Sigma \cup \overline{\Sigma})^*$,

$$g(a) = ah_1(a)\overline{h_2(a)} \; \forall a \in \Sigma_2$$

ja kieli $R_1 = g(R) \sqcup \overline{\Sigma_2}^*$, joka lauseen 1.6 nojalla on säännöllinen.

Osoitetaan, että

$$E(h_1, h_2) \cap R = pr_{\Sigma_2}(TS_\Sigma \cap R_1). \tag{51}$$

Olkoon $w \in pr_{\Sigma_2}(TS_\Sigma \cap R_1)$, tarkemmin

$$\begin{aligned} w &\in pr_{\Sigma_2}(g(r) \sqcup \overline{w_2}) \\ &= pr_{\Sigma_2}(r_1 h_1(r_1)\overline{h_2(r_1)} \cdots r_n h_1(r_n)\overline{h_2(r_n)} \sqcup \overline{w_2}),\end{aligned}$$

missä

$$r = r_1 \cdots r_n \in R, r_i \in \Sigma_2 \forall 1 \leq i \leq n, \quad \overline{w_2} \in \overline{\Sigma_2}^*.$$

Koska

$$h_1(r_i), h_2(r_i) \in \Sigma_1^*, \; r_i \in \Sigma_2 \quad \forall 1 \leq i \leq n$$

ja

$$pr_{\Sigma_2}(g(r) \sqcup \overline{w_2}) = \{r\}$$

eli $w = r$, niin yhtälöiden 50 ja 37 nojalla

$$w \in pr_{\Sigma_2}(TS_\Sigma \cap R_1) \quad \Leftrightarrow \quad w \in E(h_1, h_2) \cap R.$$

Koska $T \subseteq \Sigma_2$,

$$\begin{aligned} L &= pr_T(E(h_1, h_2) \cap R) \\ &= pr_T(pr_{\Sigma_2}(TS_\Sigma \cap R_1)) \\ &= pr_T(TS_\Sigma \cap R_1).\end{aligned}$$

Olkoot $\Sigma = \{a_1, \ldots, a_n\}$ ja $\overline{\Sigma} = \{\overline{a_1}, \ldots, \overline{a_n}\}$ aakkostoja ja $h : \{a_1, \ldots, a_n, \overline{a_1}, \ldots, \overline{a_n}\}^* \to \{0, 1, \overline{0}, \overline{1}\}^*$ morfismi

$$h(x) = \begin{cases} 01^i0 & \text{kun} \quad x = a_i, \quad 1 \leq i \leq n \\ \overline{01}^i\overline{0} & \text{kun} \quad x = \overline{a_i}, \quad 1 \leq i \leq n. \end{cases}$$

Jonokoneen

$$M = (\{q, q_0, q_1, q_2, \ldots, q_n, \overline{q_0}, \overline{q_1}, \overline{q_2}, \ldots, \overline{q_n}\}, \{0, 1, \overline{0}, \overline{1}\}, \{a_1, \ldots, a_n, \overline{a_1}, \ldots, \overline{a_n}\}, \delta, \gamma, \{q\}, \{q\}),$$

jolla on siirtymät ja tulosteet

$$\begin{array}{lcllcl}
\delta(q, 0) & = & q_0 & \gamma(q, 0) & = & \lambda \\
\delta(q_i, 1) & = & q_{i+1}, i = 0, 1, \ldots, n-1 & \gamma(q_i, 1) & = & \lambda, i = 0, 1, \ldots, n-1 \\
\delta(q_i, 0) & = & q, i = 1, 2, \ldots, n & \gamma(q_i, 0) & = & a_i, i = 1, 2, \ldots, n \\
\delta(q, \overline{0}) & = & \overline{q_0} & \gamma(q, \overline{0}) & = & \lambda \\
\delta(\overline{q_i}, \overline{1}) & = & \overline{q_{i+1}}, i = 0, 1, \ldots, n-1 & \gamma(\overline{q_i}, \overline{1}) & = & \lambda, i = 0, 1, \ldots, n-1 \\
\delta(\overline{q_i}, \overline{0}) & = & q, i = 1, 2, \ldots, n & \gamma(\overline{q_i}, \overline{0}) & = & \overline{a_i}, i = 1, 2, \ldots, n,
\end{array}$$

määrittelemä gsm-kuvaus $g_1 : \{0, 1, \overline{0}, \overline{1}\}^* \to (\Sigma \cup \overline{\Sigma})^*$ toteuttaa

$$g_1(TS_{\{0,1\}}) = h^{-1}(TS_{\{0,1\}}) = TS_\Sigma.$$

Esimerkin 1.13 nojalla on olemassa gsm-kuvaus $g_2 : (\Sigma \cup \overline{\Sigma})^* \to (\Sigma \cup \overline{\Sigma})^*$,

$$g_2(TS_\Sigma) = TS_\Sigma \cap R_1.$$

Esimerkin 1.12 nojalla on olemassa gsm-kuvaus $g_3 : (\Sigma \cup \overline{\Sigma})^* \to T^*$, missä $T \subseteq \Sigma$,

$$g_3(TS_\Sigma \cap R_1) = pr_T(TS_\Sigma \cap R_1) = L.$$

Lauseen 1.9 nojalla on olemassa gsm-kuvaus $g : \{0, 1, \overline{0}, \overline{1}\} \to T^*$,

$$g = g_3 \circ g_2 \circ g_1,$$

joka toteuttaa $L = g(TS_{\{0,1\}})$. □

Seuraavan lauseen todistuksessa on käytetty lähteenä kirjaa [12].

Lause 2.11. *Jokaista luokan RE kieltä $L \subseteq T^*$ kohti on olemassa sellainen gsm-kuvaus g, että $L = g(RTS_{\{0,1\}})$.*

Todistus. Määritellään morfismi $h : \{0, 1, \overline{0}, \overline{1}\}^* \to \{0, 1, \overline{0}, \overline{1}\}^*$,

$$\begin{array}{ll}
h(0) = 01 & h(1) = 10 \\
h(\overline{0}) = \overline{01} & h(\overline{1}) = \overline{10},
\end{array}$$

ja säännöllinen kieli

$$L' = \{01, 10, \overline{01}, \overline{10}\}^* 00\overline{00}\{01\overline{01}, 10\overline{10}\}^* \subseteq \{0, 1, \overline{0}, \overline{1}\}^*.$$

Olkoon $w = h(u)00\overline{00}v \in RTS_{\{0,1\}} \cap L'$. Näytetään, että $u \in TS_{\{0,1\}}$. Yhtälön 38 nojalla

$$pr_{\{0,1\}}(w) = \overline{pr_{\{\overline{0},\overline{1}\}}(w)}^R \tag{52}$$

ja aakkoston $\{01\overline{01}, 10\overline{10}\}$ muodon perusteella

$$pr_{\{0,1\}}(v) = \overline{pr_{\{\overline{0},\overline{1}\}}(v)^R}. \tag{53}$$

Tällöin

$$\begin{aligned} pr_{\{0,1\}}(w) &= pr_{\{0,1\}}(h(u)00\overline{00}v) \\ &= pr_{\{0,1\}}(h(u))00pr_{\{0,1\}}(v) \end{aligned} \tag{54}$$

ja peilikuvaoperaation ominaisuuden 2 ja yhtälön 53 nojalla

$$\begin{aligned} \overline{pr_{\{\overline{0},\overline{1}\}}(w)^R} &= \overline{pr_{\{\overline{0},\overline{1}\}}(h(u)00\overline{00}v)^R} \\ &= (pr_{\{\overline{0},\overline{1}\}}(h(u))\overline{00}pr_{\{\overline{0},\overline{1}\}}(v))^R \\ &= \overline{pr_{\{\overline{0},\overline{1}\}}(v)^R\overline{00}pr_{\{\overline{0},\overline{1}\}}(h(u))^R} \\ &= pr_{\{0,1\}}(v)^R00\overline{pr_{\{\overline{0},\overline{1}\}}(h(u))}^R. \end{aligned} \tag{55}$$

Yhtälön 52 nojalla sanat 54 ja 55 ovat yhtäsuuret. Ne sisältävät vain yhden parillisessa kohdassa olevan osasanan 00, jonka kirjainten indeksit ovat muotoa $2i-1, 2i$. Vertailemalla sanojen 54 ja 55 etuliitteitä parillisessa kohdassa olevaan sanaan 00 asti ja sanojen 54 ja 55 loppuliitteitä parillisessa kohdassa olevan sanan 00 jälkeen nähdään, että

$$\begin{aligned} pr_{\{0,1\}}(h(u)) &= pr_{\{0,1\}}(v)^R \\ &= \overline{pr_{\{\overline{0},\overline{1}\}}(h(u))}. \end{aligned}$$

Koska morfismi h määrittelystään johtuen kommutoi projektiokuvausten $pr_{\{0,1\}}$ ja $pr_{\{\overline{0},\overline{1}\}}$ ja morfismin c kanssa,

$$h(pr_{\{0,1\}}(u)) = h(\overline{pr_{\{\overline{0},\overline{1}\}}(u)}). \tag{56}$$

Koska h on injektio,

$$pr_{\{0,1\}}(u) = \overline{pr_{\{\overline{0},\overline{1}\}}(u)}$$

eli yhtälön 37 nojalla

$$u \in TS_{\{0,1\}}.$$

Esimerkin 1.13 nojalla on olemassa gsm-kuvaus $g_1 : \{0,1,\overline{0},\overline{1}\}^* \to \{0,1,\overline{0},\overline{1}\}^*$, joka toteuttaa

$$g_1(RTS_{\{0,1\}}) = RTS_{\{0,1\}} \cap L'.$$

Jonokoneen

$$M = (\{q, q_0, q_1, \overline{q_0}, \overline{q_1}, q_f\}, \{0, 1, \overline{0}, \overline{1}\}, \{0, 1, \overline{0}, \overline{1}\}, \delta, \gamma, \{q\}, \{q_f\}),$$

jolla on siirtymät ja tulosteet

$$\begin{array}{lcllcl}
\delta(q,i) & = & q_i, i=0,1 & \gamma(q,i) & = & \lambda, i=0,1 \\
\delta(q,\overline{i}) & = & \overline{q_i}, i=0,1 & \gamma(q,\overline{i}) & = & \lambda, i=0,1 \\
\delta(q_0,1) & = & q & \gamma(q_0,1) & = & 0 \\
\delta(q_1,0) & = & q & \gamma(q_1,0) & = & 1 \\
\delta(\overline{q_0},\overline{1}) & = & q & \gamma(\overline{q_0},\overline{1}) & = & \overline{0} \\
\delta(\overline{q_1},\overline{0}) & = & q & \gamma(\overline{q_1},\overline{0}) & = & \overline{1} \\
\delta(q_0,0) & = & q_f & \gamma(q_0,0) & = & \lambda \\
\delta(q_f,i) & = & q_f, i=0,1,\overline{0},\overline{1} & \gamma(q_f,i) & = & \lambda, i=0,1,\overline{0},\overline{1},
\end{array}$$

määrittelemä gsm-kuvaus $g_2 : \{0, 1, \overline{0}, \overline{1}\}^* \to \{0, 1, \overline{0}, \overline{1}\}^*$, toteuttaa

$$g_2(h(u)00\overline{00}v) = u.$$

Näin ollen lauseen 1.9 nojalla on olemassa gsm-kuvaus $g : \{0, 1, \overline{0}, \overline{1}\}^* \to \{0, 1, \overline{0}, \overline{1}\}^*$, joka on gsm-kuvausten g_2 ja g_1 kompositio ja joka toteuttaa $g(RTS_{\{0,1\}}) \subseteq TS_{\{0,1\}}$.

Olkoon $w \in TS_{\{0,1\}}$. Määritellään sana

$$w' = h(w)00\overline{00}v \in L',$$

jonka loppuliite $v \in \{01\overline{01}, 10\overline{10}\}^*$ määräytyy ehdosta

$$pr_{\{0,1\}}(v) = pr_{\{0,1\}}(h(w))^R. \tag{57}$$

Sana v on olemassa ja yksikäsitteinen. Morfismin h määrittelyn ja aakkoston $\{01\overline{01}, 10\overline{10}\}$ muodon perusteella sanalle w' pätee

$$\begin{aligned}
\overline{(pr_{\{0,1\}}(w'))^R} &= \overline{(pr_{\{0,1\}}(h(w)00\overline{00}v))^R} \\
&= \overline{(pr_{\{0,1\}}(h(w))00pr_{\{0,1\}}(v))^R} \\
&= \overline{(pr_{\{0,1\}}(v))^R00(pr_{\{0,1\}}h(w))^R} \\
&= \overline{(pr_{\{0,1\}}(v))^R00pr_{\{0,1\}}(v)} \\
&= \overline{pr_{\{0,1\}}h(w)00pr_{\{0,1\}}(v)} \\
&= pr_{\{\overline{0},\overline{1}\}}h(w)\overline{00}pr_{\{\overline{0},\overline{1}\}}(v) \\
&= pr_{\{\overline{0},\overline{1}\}}(h(w)00\overline{00}v) \\
&= pr_{\{\overline{0},\overline{1}\}}(w'),
\end{aligned}$$

joten yhtälön 38 nojalla

$$w' \in RTS_{\{0,1\}}.$$

Jokaista kieleen $TS_{\{0,1\}}$ kuuluvaa sanaa w kohti on siis olemassa kieleen $RTS_{\{0,1\}} \cap L'$ kuuluva sana w', jonka gsm-kuvaus g kuvaa sanaksi w eli $TS_{\{0,1\}} \subseteq g(RTS_{\{0,1\}})$. Väite seuraa nyt lauseesta 2.10. □

3 H-systeemit

3.1 Silmukointioperaatio

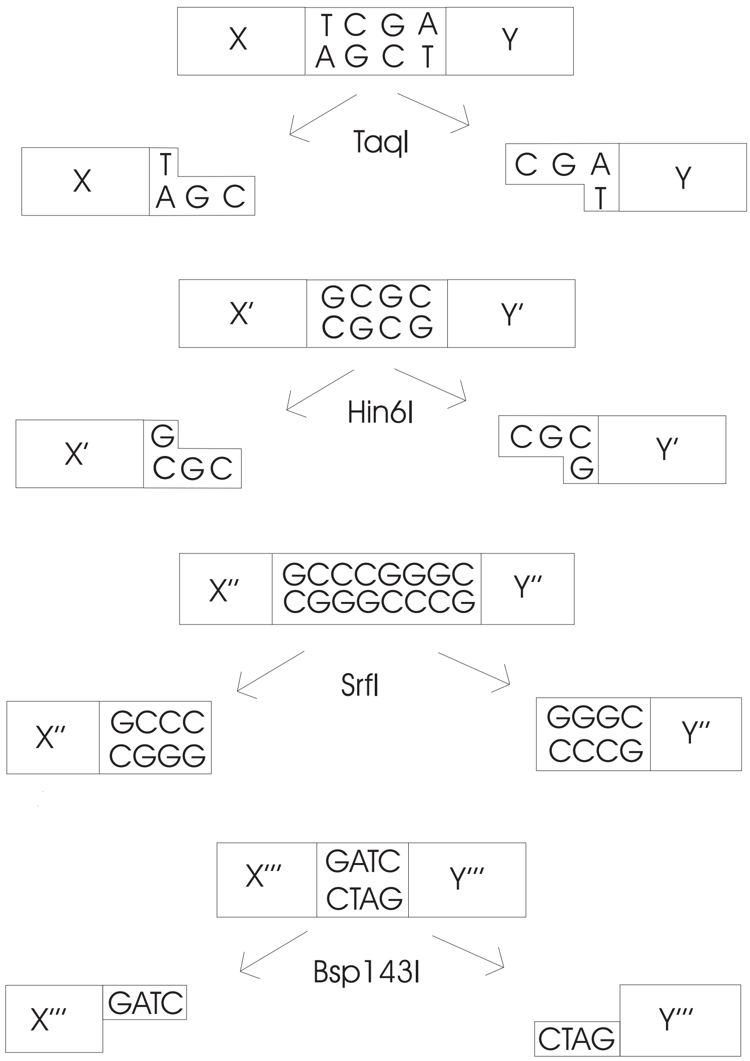
Entsyymit ovat proteiineja, jotka mahdollistavat tietyn biokemiallisen reaktion tapahtumisen säilyen itse muuttumattomana. *Ligaasi* on entsyymi, joka mahdollistaa säikeiden liittymisen kaksoiskierteeksi. *Restriktioentsyymit* ovat erityisesti bakteereissa tavattavia entsyymejä, jotka pilkkovat DNA-kaksoiskierteitä.

Restriktioentsyymien toiminnan kontrolloinnin kannalta hyödyllisimpiä ovat sellaiset restriktioentsyymit, jotka pilkkovat vain kyseiselle restriktioentsyymille ominaisen kaksoiskierteen sisältäviä kaksoiskierteitä. Tätä kaksoiskierrettä $(w, \overline{w^R})$ sanotaan restriktioentsyymin *tunnisteeksi*. On olemassa restriktioentsyymejä, joiden tunniste sisältää monikäsitteisyyttä, mutta jatkossa tarkastellaan vain sellaisia restriktioentsyymejä, joiden tunniste on yksikäsitteinen. Lisäksi suljetaan pois tarkasteluista sellaiset restriktioentsyymit, joilla pilkkominen tapahtuu tunnisteen ulkopuolella eli sellaiset, joilla tunniste säilyy ehjänä.

Esimerkki 3.1. Restriktioentsyymit $TaqI$, $Hin6I$, $SrfI$ ja $Bsp143I$, joilla on tunnisteet $(TCGA, TCGA)$, $(GCGC, GCGC)$, $(GCCCGGGC, GCCCGGGC)$ ja $(GATC, GATC)$, pilkkovat tunnisteensa sisältävät kaksoiskierteet kuvassa 2 esitetyllä tavalla.

Jatkossa kutsutaan restriktioentsyymin r pilkkomisen tuloksena syntyneitä molekyylejä restriktioentsyymin *r tuottamiksi* molekyyleiksi. Kuten kuvasta 2 nähdään, restriktioentsyymien $TaqI$ ja $Hin6I$ tuottamilla molekyyleillä on lyhyt kaksoiskierteen yli ulottuva säie, jota kutsutaan *koheesiiviseksi pääksi*. $SrfI$ on esimerkki restriktioentsyymistä, jonka tuottamilla molekyyleillä ei ole koheesiivista päätä. $Bsp143I$ on esimerkki restriktioentsyymistä, jonka tuottamien molekyylien koheesiivinen pää on koko tunnisteen pituinen.

Seuraavaan taulukkoon on koottu luettelosta [11] joitakin kaupallisesti saatavia restriktioentsyymejä, joiden tuottamilla molekyyleillä on koheesiivinen pää. On erittäin yleistä, että yksikäsitteinen tunniste on muotoa (w, w), jolloin sana w on biologinen palindromi. Lisäksi koheesiiviset päät sijaitsevat



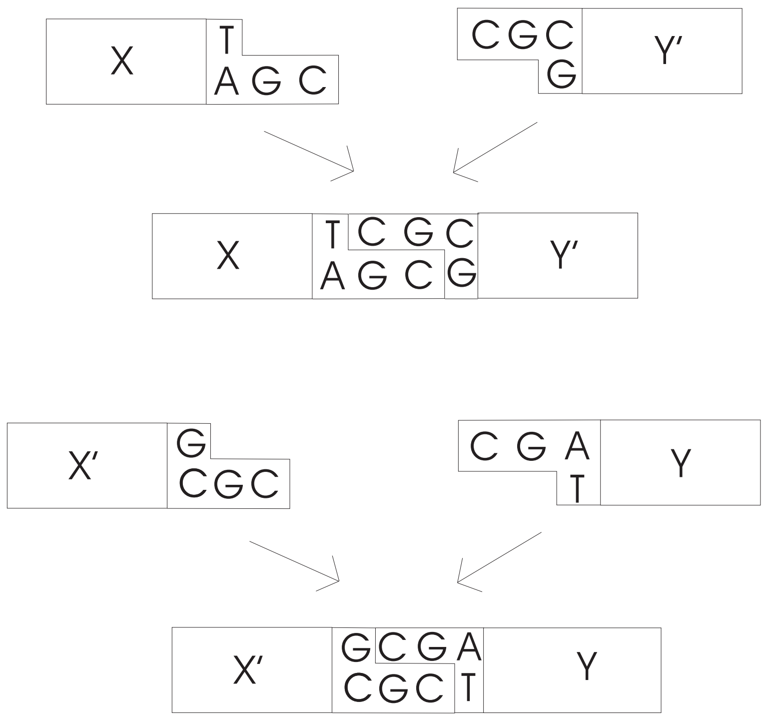
Kuva 2: Restriktioentsyymien *TaqI*, *Hin6I*, *SrfI* ja *Bsp143I* toiminta

tunnisteessa keskellä, joten restriktioentsyymin, jonka tunniste on muotoa (w, w), tuottamat koheesiiviset päät ovat biologisia palindromeja. Näin ollen, koska taulukon kaikkien restriktioentsyymien tunnisteet ovat kyseistä muotoa, niiden tuottamat koheesiiviset päät on voitu esittää yhdellä sanalla w, joka on saatu luettelemalla koheesiivisen pään nukleotidit suuntaan 5'-3'. Tällöin on otettu huomioon se, että restriktioentsyymit jakautuvat kahteen luokkaan sen mukaan, päättyvätkö niiden tuottamat koheesiiviset päät 5'- vai 3'-nukleotideihin. 5'-nukleotidejä tuottavat restriktioentsyymit ovat taulukon yläosassa ja 3'-nukleotidejä tuottavat taulukon alaosassa. Koheesiiviset päät on erotettu tunnisteiden sisällä pystyviivoin.

restriktioentsyymi	tunniste	w, $\lvert w\rvert = 2$	w, $\lvert w\rvert = 4$
$Psp1406I$	$(AA\vert CG\vert TT, AA\vert CG\vert TT)$	CG	
$TasI$	$(\vert AATT\vert, \vert AATT\vert)$		$AATT$
$BspLU11I$	$(A\vert CATG\vert T, A\vert CATG\vert T)$		$CATG$
$Bsu15I$	$(AT\vert CG\vert AT, AT\vert CG\vert AT)$	CG	
$VspI$	$(AT\vert TA\vert AT, AT\vert TA\vert AT)$	TA	
$NdeI$	$(CA\vert TA\vert TG, CA\vert TA\vert TG)$	TA	
$NcoI$	$(C\vert CATG\vert G, C\vert CATG\vert G)$		$CATG$
$MaeI$	$(C\vert TA\vert G, C\vert TA\vert G)$	TA	
$Hin6I$	$(G\vert CG\vert C, G\vert CG\vert C)$	CG	
$NotI$	$(GC\vert GGCC\vert GC, GC\vert GGCCGC)$		$GGCC$
$NheI$	$(G\vert CTAG\vert C, G\vert CTAG\vert C)$		$CTAG$
$TaqI$	$(T\vert CG\vert A, T\vert CG\vert A)$	CG	
$TaiI$	$(\vert ACGT\vert, \vert ACGT\vert)$		$ACGT$
$Mph1103I$	$(A\vert TGCA\vert T, A\vert TGCA\vert T)$		$TGCA$
$NlaIII$	$(\vert CATG\vert, \vert CATG\vert)$		$CATG$
$SdaI$	$(CC\vert TGCA\vert GG, CC\vert TGCA\vert GG)$		$TGCA$
$AatII$	$(G\vert ACGT\vert C, G\vert ACGT\vert C)$		$ACGT$
$PaeI$	$(G\vert CATG\vert C, G\vert CATG\vert C)$		$CATG$
$Cfr42I$	$(CC\vert GC\vert GG, CC\vert GC\vert GG)$	GC	

Seuraavaksi esiteltävä tapahtuma on ollut lähtökohtana myöhemmin tässä luvussa määriteltäville laskentamalleille. Restriktioentsyymien $TaqI$ ja $Hin6I$ tuottamien molekyylien koheesiiviset päät voidaan asettaa kohdakkain siten, että kaikkien koheesiivisten päiden nukleotidien kohdalla on kyseisen nukleotidin komplementti. Lisäksi restriktioentsyymien $TaqI$ ja $Hin6I$ tuottamilla koheesiivisilla päillä on 5'-nukleotidi siinä päässä, jolla on vain yksi viereinen nukleotidi. Tällöin ligaasin läsnäollessa restriktioentsyymien

TaqI ja *Hin6I* tuottamat molekyylit voivat yhdistyä ristiin kuvassa 3 esitetyllä tavalla. Koska molemmilla koheesiivilla päillä on 5'-nukleotidi, ne liittyvät 3'-nukleotideihin, mikä on ainoa tapa, jolla vierekkäiset nukleotidit voivat liittyä toisiinsa.



Kuva 3: Restriktioentsyymien *TaqI* ja *Hin6I* tuottaminen molekyylien ristiinyhdistymien

Restriktioentsyymien *TaqI* ja *Hin6I* tuottamat molekyylit voivat ligaasin läsnäollessa luonnollisesti yhdistyä alkuperäisiksi molekyyleiksi, mutta ristiin yhdistyminen on kiinnostavampi tapaus, koska tällöin voi syntyä uusia molekyylejä.

Olkoot restriktioentsyymin r_1 tuottamat koheesiiviset päät w_1 ja restriktioentsyymin r_2 tuottamat koheesiiviset päät w_2, jolloin rajoitutaan restriktioentsyymeihin, joiden tuottamat koheesiiviset päät ovat biologisia palindromeja. Jatkossa kutsutaan restriktioentsyymejä r_1 ja r_2 *yhteensopiviksi*, jos

(i) $w_1 = w_2$ ja

(ii) koheesiivisilla päillä w_1, w_2 on joko 5'- tai 3'-nukleotidit.

Jos molemmilla ristiinyhdistymiseen osallistuvilla koheesiivisilla päillä on 3'-nukleotidit, niin ristiinyhdistymisessä koheesiivisten päiden 3'-nukleotidit liittyvät 5'-nukleotideihin. Vastaavasti jos molemmilla koheesiivisilla päillä on 5'-nukleotidit, niin ne liittyvät 3'-nukleotideihin. Edellä esitetyn taulukon avulla voidaan helposti löytää yhteensopivia restriktioentsyymejä valitsemalla taulukon joko ylä- tai alaosasta restriktioentsyymit, jotka tuottavat samat koheesiiviset päät.

Määritellään sanojen välinen operaatio, jolla voidaan mallintaa yhteensopivien restriktioentsyymien tuottamien molekyylien ristiinyhdistymistä. Olkoon V aakkosto ja $\#$ ja $\$$ sen ulkopuolisia symboleja. *Silmukointisääntö* on kieleen

$$V^*\#V^*\$V^*\#V^*$$

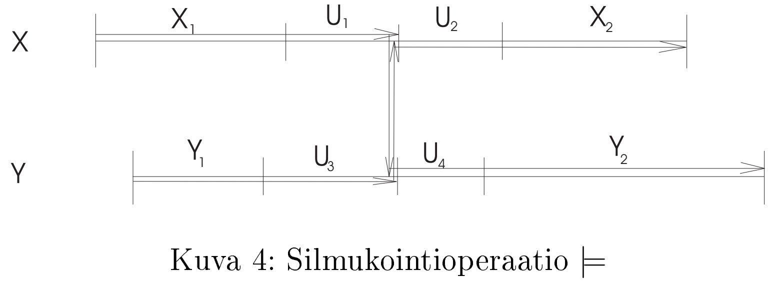
kuuluva sana, joka myöhemmin määriteltävissä malleissa vastaa yhteensopivien restriktioentsyymien paria. Silmukointisääntö $u_1\#u_2\$u_3\#u_4$ voidaan esittää ristikkomuodossa

$$\begin{array}{c|c} u_1 & u_2 \\ \hline u_3 & u_4. \end{array}$$

Silmukointioperaatio $\models$ on kahden sanan $x, y \in V^*$ välinen operaatio, jonka tuloksena on kaksi sanaa $z, w \in V^*$ ja jossa *silmukoidaan* eli sovelletaan tiettyä *silmukointisääntöä* $r \in V^*\#V^*\$V^*\#V^*$:

$$(x, y) \models_r (z, w) \quad \Leftrightarrow \quad \begin{cases} x = x_1u_1u_2x_2, & z = x_1u_1u_4y_2 \\ y = y_1u_3u_4y_2, & w = y_1u_3u_2x_2 \\ r = u_1\#u_2\$u_3\#u_4. \end{cases}$$

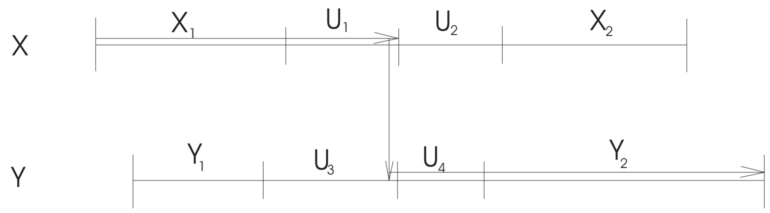
Silmukointioperaatiossa $\models$ sanat x ja y vastaavat kaksoiskierteitä, joista kumpikin sisältää yhteensopivista restriktioentsyymeistä jommankumman tunnisteen ja sanat z ja w kyseisten restriktioentsyymien pilkkomisen ja ristiinyhdistymisen tuloksena syntyviä molekyylejä. Edellytyksenä sille, että silmukointisääntöä $u_1\#u_2\$u_3\#u_4$ voidaan soveltaa sanoihin x ja y on, että sana u_1u_2 on sanan x osasana ja että sana u_3u_4 on sanan y osasana. Silmukoinnin tuloksena saadut sanat z ja w saadaan katenoimalla sanan x etuliite x_1u_1 sanan y loppuliitteen u_4y_2 kanssa ja katenoimalla sanan y etuliite y_1u_3 sanan x loppuliitteen x_1u_1 kanssa.



Kuva 4: Silmukointioperaatio $\models$

Seuraavaksi määriteltävä silmukointioperaatio $\vdash$ eroaa silmukointioperaatiosta $\models$ vain siten, että se antaa tulokseksi yhden sanan. Myöhemmissä tarkasteluissa kaikki silmukointia koskevat tulokset todistetaan silmukointioperaatiolle $\models$, käyttäen silmukointioperaatiota $\vdash$ vain apuoperaationa, koska restriktioentsyymien pilkkomien molekyylien ristiinyhdistymisessä molemmat ristiinyhdistymistavat ovat yhtä todennäköisiä. Silmukointioperaatio $\vdash$ määräytyy ehdosta

$$(x, y) \vdash_r z \quad \Leftrightarrow \quad \begin{cases} x \;=\; x_1u_1u_2x_2, & z \;=\; x_1u_1u_4y_2 \\ y \;=\; y_1u_3u_4y_2, & r \;=\; u_1\#u_2\$u_3\#u_4. \end{cases}$$



Kuva 5: Silmukointioperaatio $\vdash$

3.2 Laajennettujen H-systeemien generointivoima

Määritellään silmukointioperaation $\models$ avulla H-systeemeiksi kutsuttu kielen generointimekanismi, jolla voidaan mallintaa kaksoiskierteiden joukolla operoivia yhteensopivien restriktioentsyymien pareja ligaasin läsnäollessa. Mallin taustaoletuksena on, että sekä alkuperäisiä molekyylejä että jokaista ristiinyhdistymiseen tuloksena syntyvää molekyylejä on aina tarvittaessa saatavilla. Saatavuudesta tulisi huolehtia käyttäen monistustekniikkoja.

Laajennetun H-systeemin $\gamma = (N, T, A, R)$ komponenteista N ja T ovat erilliset väli- ja pääteaakkostot, joiden unionista käytetään merkintää $V = N \cup T$, $A \subset V^*$ on *aksioomien* joukko ja

$$R \subset V^*\#V^*\$V^*\#V^*$$

on äärellinen silmukointisääntöjen joukko. Jos $N = \emptyset$, kysymyksessä on *laajentamaton* H-systeemi. Kuvauksen $\sigma : \mathcal{P}(V^*) \to \mathcal{P}(V^*)$,

$$\sigma(L) = \{z \in V^* \mid (x, y) \vdash_r (z, w) \text{ tai } (x, y) \vdash_r (w, z) \text{ joillakin } x, y \in V^*, r \in R\},$$

jonka iteroinnista käytetään merkintöjä

$$\sigma^0(L) \;=\; L$$

$$\begin{aligned}\sigma^{i+1}(L) &= \sigma^i(L) \cup \sigma(\sigma^i(L))\\ \sigma^*(L) &= \bigcup_{i\geq 0} \sigma^i(L),\end{aligned}$$

avulla voidaan määritellä laajennetun H-systeemin γ *generoima* kieli

$$L(\gamma) = \sigma^*(A) \cap T^*.$$

Se koostuu niistä sanoista yli pääteaakkoston, jotka voidaan saada soveltamalla aksioomiin ja silmukoinnin tuloksena saatuihin sanoihin silmukointioperaatiota $\vdash$ äärellisen monta kertaa.

Huomautus 3.2. Leikkaus kielen T^* kanssa voidaan toteuttaa biokemiallisesti suodattamalla.

Esimerkki 3.3. Olkoon $\gamma = (\{b, c\}, \{a\}, \{a^n b, ca^m\}, \{a\#b\$c\#a\})$, missä n ja m ovat positiivisia kokonaislukuja, laajennettu H-systeemi. Silmukointisääntöä $a\#b\$c\#a$ voi soveltaa vain sanoihin, joista toinen sisältää sanan ab ja toinen sanan ca osasananaan, joten ainoa mahdollisuus alussa on soveltaa silmukointisääntöä aksioomiin, jolloin saadaan

$$(a^n|b, c|a^m) \models (a^{n+m}, cb),$$

missä silmukointikohdat on osoitettu pystyviivalla. Koska kumpikaan sanoista a^{n+m} ja cb ei sisällä sanaa ab tai ca osasananaan, sanat a^{n+m} ja cb eivät voi esiintyä H-systeemin γ silmukoinnissa, joten

$$\sigma^*(A) = \sigma^0(A) \cup \sigma^1(A) = \{a^n b, ca^m, a^{n+m}, cb\}$$

ja

$$L(\gamma) = \sigma^*(A) \cap \{a\}^* = \{a^{n+m}\}.$$

H-systeemi γ poislukien sen aksioomat on kiinteä mekanismi, joka laskee kahden aksioomissa annetun luonnollisen luvun summan.

Seuraavaksi määriteltävää mekanismia, joka poikkeaa laajennetuista H-systeemeistä vain silmukointioperaation $\vdash$ sijasta käytetyn silmukointioperaation $\models$ osalta, tarvitaan vain lauseen 3.4 todistuksen yksinkertaistamista varten. *Laajennetun $H_\vdash$-systeemin* $\gamma_\vdash = (N, T, A, R)$ komponentit ovat kuin laajennetuilla H-systeemeillä ja jos $N = \emptyset$, kysymyksessä on *laajentamaton* $H_\vdash$-systeemi. Kuvauksen $\sigma_\vdash : \mathcal{P}(V^*) \to \mathcal{P}(V^*)$,

$$\sigma_\vdash(L) = \{z \in V^* \mid (x, y) \vdash_r z \text{ joillakin } x, y \in L, r \in R\},$$

jota voidaan iteroida vastaavasti kuin kuvausta σ, avulla voidaan määritellä laajennetun $H_\vdash$-systeemin $\gamma_\vdash$ *generoima* kieli

$$L(\gamma_\vdash) = \sigma_\vdash^*(A) \cap T^*.$$

Seuraavan lauseen, jonka todistivat ensimmäisenä Culik ja Harju [4], todistuksessa on käytetty lähteenä artikkelia [13].

Lause 3.4. *Jos $\gamma_\vdash = (\emptyset, T, A, R)$ on laajentamaton $\gamma_\vdash$-systeemi, jonka aksioomien joukko on säännöllinen kieli, niin $L(\gamma_\vdash)$ on säännöllinen.*

Todistus. Olkoon $\gamma_\vdash = (\emptyset, T, A, R)$ laajentamaton $H_\vdash$-systeemi, jolle pätee $A \in Reg$ ja olkoon $M_0 = (Q, T, \delta, Q_0, F)$ yleistetty äärellinen automaatti, jonka tunnistama kieli on A. Muodostetaan jokaista joukon R silmukointisääntöä $r = u_1\#u_2\$u_3\#u_4$ kohti suunnattu leimattu graafi

$$B(r) = (\{i(r), i_1, i_2, \ldots, i_r, t(r)\}, \{(i(r), i_1), (i_1, i_2), \ldots, (i_r, t(r))\}, \epsilon_r),$$

jonka kukin viiva on leimattu yhdellä kirjaimella siten, että

$$\epsilon((i(r), i_1)(i_1, i_2) \cdots (i_r, t(r)) = u_1u_4.$$

Sanasta u_1u_4 käytetään merkintää $b(r)$. Graafien $B(r_i)$ pisteiden joukot ovat erillisiä. Olkoon B graafi, jonka pisteet ja viivat ovat graafien $B(r_i)$ pisteiden ja viivojen unioni ja jonka leimafunktion ϵ arvot määräytyvät graafien $B(r_i)$ leimafunktioiden arvoista. Olkoon G_0 graafi, joka on saatu automaatin M_0 esityksestä graafina lisäämällä siihen graafin B pisteet ja viivat leimoineen. Tällöin graafin G_0 hyväksyvät polut ovat tarkalleen ne, jotka on leimattu kielen $L(M_0)$ sanoilla, koska joukon Q pisteistä ei ole viivoja graafista B peräisin oleviin pisteisiin.

Muodostetaan sellainen yleistettyjen äärellisten automaattien jono M_i, missä $i \geq 0$, jonka jäsenet toteuttavat ehdon

$$\sigma_\vdash(L(M_k)) \subseteq L(M_{k+1}), k \geq 0. \tag{58}$$

Kaikilla automaateilla M_k tulee olemaan samat alku- ja lopputilojen joukot, joten automaatit M_k voidaan esittää muotoa

$$G_k = ((Q_k, E_k, \epsilon_k), Q_0, F)$$

olevina graafeina. Seuraava konstruktio antaa automaatista M_k automaatin M_{k+1}, missä $k \geq 0$. Tarkastellaan niitä joukon R silmukointisääntöjä $r = u_1\#u_2\$u_3\#u_4$, joita kohti on olemassa sellaiset kieleen $L(M_k)$ kuuluvat sanat w ja w', että

$$u_1u_2 \in F(w), \quad u_3u_4 \in F(w'),$$

ja kaikkia graafin G_k hyväksyviä polkuja. Jos

(i) pqt on hyväksyvä polku, jossa polku q alkaa pisteestä v,

(ii) $\epsilon_k(q) = u_1u_2$ ja

(iii) graafissa G_k ei ole viivaa pisteestä v pisteeseen $i(r)$,

niin lisätään graafiin G_k sanalla λ leimattu viiva

$$(v, i(r)). \tag{59}$$

Jos

(i) $p'q't'$ on hyväksyvä polku, jossa polku q' päättyy pisteeseen v',

(ii) $\epsilon_k(q') = u_3u_4$ ja

(iii) graafissa G_k ei ole viivaa $(t(r), v')$,

lisätään graafiin G_k sanalla λ leimattu viiva

$$(t(r), v'). \tag{60}$$

Olkoon G_{k+1} prosessin tuloksena saatu graafi, joka määrää yksikäsitteisen yleistetyn äärellisen automaatin M_{k+1}. Kutsutaan graafiin G_k lisättyjä viivoja 59 *tyypin* $k+1$ *alkumerkeiksi* ja viivoja 60 *tyypin* $k+1$ *loppumerkeiksi* ja merkitään niitä aakkoston T ulkopuolisilla kirjaimilla $\{$ ja $\}$. Graafi G_{k+1} saadaan siis graafista G_k lisäämällä graafin G_k viivoihin sanalla λ leimatut tyypin $k+1$ alku- ja loppumerkit. Tällöin konstruktion nojalla ehto 58 toteutuu.

Koska konstruktiossa lisätään ainoastaan viivoja pisteiden pysyessä ennallaan, alku- ja loppumerkkien lukumäärä on ylhäältä rajoitettu. On siis olemassa sellainen indeksi n, että

$$G_{n+k} = G_n \quad \forall k \geq 1.$$

Näytetään, että kieli $L(M_n)$ on suljettu silmukointioperaation $\vdash$ suhteen, mistä seuraa, että

$$\begin{aligned} L(\gamma_\vdash) &= \sigma_\vdash^*(A) \\ &= \sigma_\vdash^*(L(M_0)) \\ &= \bigcup_{i\geq 0} \sigma_\vdash^i(L(M_0)) \\ &\subseteq \bigcup_{i\geq 0} L(M_i) \\ &= L(M_n). \end{aligned}$$

Olkoot

$$w = w_1 u_1 u_2 w_2, \quad w' = w_1' u_3 u_4 w_2'$$

kieleen $L(M_n)$ kuuluvia sanoja ja $r = u_1 \# u_2 \$ u_3 \# u_4 \in R$ silmukointisääntö. Olkoot pqt ja $p'q't'$ sellaiset graafin G_n hyväksyvät polut, että

$$\begin{array}{lll} \epsilon_n(p) = w_1 & \epsilon_n(q) = u_1 u_2 & \epsilon_n(t) = w_2 \\ \epsilon_n(p') = w_1' & \epsilon_n(q') = u_3 u_4 & \epsilon_n(t') = w_2', \end{array}$$

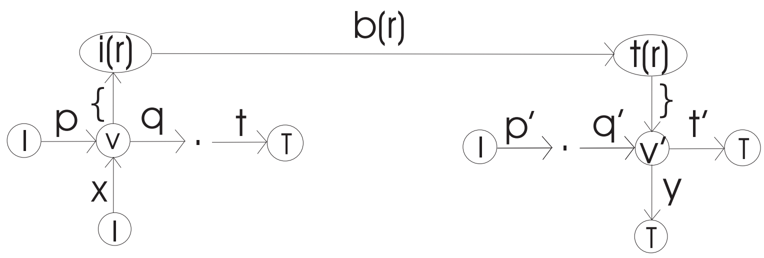
viivan q alkupiste on v ja viivan q' päätepiste on v'. Tällöin pisteestä v on alkumerkki $\{$ pisteeseen $i(r)$ ja pisteestä $t(r)$ on loppumerkki $\}$ pisteeseen v'. Näin ollen polku

$$p\{b(r)\}t'$$

on graafin G_n hyväksyvä polku ja

$$\epsilon_n(p\{b(r)\}t') = w_1 u_1 u_4 w_2',$$

joten sanojen w ja w' silmukoinnin tulos käyttäen silmukointisääntöä r kuuluu kieleen $L(M_n)$.



Kuva 6: Hyväksyvät polut automaatin M_n esityksessä graafina

Näytetään, että $L(M_n) \subseteq L(\gamma_\vdash)$. Liitetään jokaiseen graafin G_n polkuun w vektori $c = (c_1, c_2, \ldots, c_n) \in N^n$, jonka komponentit c_i, missä $1 \leq i \leq n$, ovat yhteenlaskettujen tyyppiä i olevien alku- ja loppumerkkien lukumäärät polussa w. Vektoria c sanotaan polun w *kompleksisuudeksi*. Määritellään kompleksisuusvektoreille c, d relaatio

$$c \prec d \quad \Leftrightarrow \quad (\exists k : 1 \leq k \leq n,\ c_k < d_k \quad \text{ja} \quad c_j = d_j,\ j = k+1, k+2, \ldots, n). \tag{61}$$

Todistetaan induktiolla polkujen kompleksisuuden suhteen, että

$$\epsilon_n(w') \in L(\gamma_\vdash)$$

kaikilla graafin G_n hyväksyvillä poluilla w'. Olkoon w' graafin G_n hyväksyvä polku. Jos polun w' kompleksisuus on $(0,\dots,0)$, niin se ei sisällä yhtään alku- tai loppumerkkiä, joten

$$w' \in A \subseteq L(\gamma_{\vdash}).$$

Olkoon polun w' kompleksisuus positiivinen ja oletetaan, että kaikki graafin G_n hyväksyvät polut v, joiden kompleksisuus on pienempi kuin c, toteuttavat ehdon

$$\epsilon_n(v) \in L(\gamma_{\vdash}).$$

Koska polun w' kompleksisuus on positiivinen ja se on hyväksyvä polku, polku w' sisältää ainakin yhden alku- ja loppumerkin. Valitaan polun w' alkumerkeistä viimeinen vasemmalta oikealle luettaessa ja kyseisen alkumerkin jälkeen ensimmäinen loppumerkeistä. Tällöin polku w' voidaan kirjoittaa muotoon

$$w' = x\{z\}y,$$

missä polku z ei sisällä alku- tai loppumerkkejä. Olkoon $\{$ tyypin k alkumerkki ja $\}$ tyypin k' loppumerkki. Olkoon $r = u_1\#u_2\$u_3\#u_4$ on silmukointisääntö, jonka yhteydessä alku- ja loppumerkit $\{$ ja $\}$ ovat lisätty graafiin G_n, jolloin $z = b(r)$. Käytetään viivan $\{$ alkupisteestä merkintää v ja viivan $\}$ päätepisteestä merkintää v'. Silloin graafissa G_{k-1} on hyväksyvä polku pqt, jossa polun q alkupiste on v ja

$$\epsilon_n(q) = u_1u_2.$$

Graafissa $G_{k'-1}$ on vastaavasti hyväksyvä polku $p'q't'$, jossa polun q' päätepiste on v' ja

$$\epsilon_n(q') = u_3u_4.$$

Polut xqt ja $p'q'y$ ovat hyväksyviä polkuja, ja niille pätee

$$(\epsilon_n(xqt), \epsilon_n(p'q'y)) \vdash_{u_1\#u_2\$u_3\#u_4} \epsilon_n(x)u_1u_4\epsilon_n(y) = \epsilon_n(x\{z\}y), \tag{62}$$

joten sana $\epsilon_n(w')$ on sanojen $\epsilon_n(xqt)$ ja $\epsilon_n(p'q'y)$ silmukoinnin tulos käyttäen silmukointisääntöä r.

Näytetään, että polkujen xqt ja $p'q'y$ kompleksisuus on pienempi kuin polun w'. Tällöin induktio-olettamuksesta seuraa, että

$$\epsilon_n(xqt), \epsilon_n(p'q'y) \in L(\gamma_{\vdash})$$

ja operaation 62 nojalla

$$\epsilon_n(w') \in L(\gamma_{\vdash}).$$

Olkoon d polun xqt kompleksisuus. Koska polku qt on graafin G_{k-1} polku, kaikki polun qt alku- ja loppumerkit ovat korkeintaan tyyppiä $k-1$. Jos siis polussa xqt on joitakin tyyppiä j, missä $j \geq k$, olevia alku- tai loppumerkkejä, niiden täytyy olla peräisin polusta x, joten ne ovat myös polussa w'. Näin ollen

$$d_j \leq c_j \; \forall j \geq k. \tag{63}$$

Polussa xqt on vähemmän tyyppiä k olevia merkkejä kuin polussa $w' = x\{z\}y$ alkumerkin $\{$ takia eli

$$d_k < c_k. \tag{64}$$

Jos yhtälö

$$d_j = c_j$$

toteutuu kaikilla indekseillä $k + 1 \leq j \leq n$, niin

$$d \prec c.$$

Jos ehto

$$d_j < c_j$$

toteutuu joillakin indekseillä j, missä $k + 1 \leq j \leq n$, niin niistä suurimman jälkeen vektorien komponentit yhtyvät, joten tällöinkin $d \prec c$.

Symmetrinen päättely osoittaa, että polun $p'q'y$ kompleksisuus on pienempi kuin polun w'. Olkoon e polun $p'q'y$ kompleksisuus. Koska $p'q'$ on graafin $G_{k'-1}$ polku, kaikki polun $p'q'$ alku- ja loppumerkit ovat korkeintaan tyyppiä $k' - 1$. Jos polussa $p'q'y$ on tyyppiä j, missä $j \geq k'$, olevia alku- tai loppumerkkejä, niiden täytyy olla peräisin polusta y, joten ne ovat myös polussa w'. Näin ollen

$$e_j \leq c_j \; \forall j \geq k'.$$

Koska polussa $p'q'y$ on vähemmän tyypin k' merkkejä kuin polussa w' loppumerkin $\}$ takia,

$$e_{k'} < c_{k'}.$$

Vastaavasti kuin edellä voidaan päätellä, että $e \prec c$. □

Seuraus 3.5. *Jos $\gamma = (\emptyset, T, A, R)$ on laajentamaton H-systeemi, jonka aksioomien joukko muodostaa säännöllisen kielen, niin $L(\gamma)$ on säännöllinen.*

Todistus. Jokaista laajentamatonta H-systeemiä $\gamma = (\emptyset, T, A, R)$, jolle pätee $A \in Reg$, kohti voidaan muodostaa laajentamaton $H_\vdash$-systeemi $\gamma'_\vdash = (\emptyset, T, A, R)$, jolla on silmukointisäännöt

$$R' = \bigcup_{u_1\#u_2\$u_3\#u_4 \in R} \{u_1\#u_2\$u_3\#u_4, u_3\#u_4\$u_1\#u_2\}.$$

Tällöin $L(\gamma) = L(\gamma'_\vdash)$ ja lauseen 3.4 nojalla $L(\gamma'_\vdash) \in Reg$. □

Seuraus 3.6. *Jos $\gamma = (N, T, A, R)$ on laajennettu H-systeemi, jonka aksioomien joukko muodostaa säännöllisen kielen, niin $L(\gamma)$ on säännöllinen.*

Todistus. Jos $\gamma = (N, T, A, R)$ on laajennettu H-systeemi, jolle pätee $A \in Reg$, niin laajentamattomalle H-systeemille

$$\gamma' = (\emptyset, N \cup T, A, R)$$

pätee seurauksen 3.5 nojalla $L(\gamma') \in Reg$. Koska lauseen 1.6 nojalla säännöllisten kielten luokka on suljettu leikkauksen suhteen ja $T^* \in Reg$, niin $L(\gamma) = L(\gamma') \cap T^* \in Reg$. □

3.3 Laajennetut mH-systeemit ja niiden universaalisuus

Seurauksen 3.6 mukaan laajennetut H-systeemit, joiden aksioomien ja silmukointisääntöjen joukot ovat äärelliset, voivat saavuttaa korkeintaan äärellisten automaattien laskentavoiman. Seuraavaksi lisätään H-systeemeihin kyky mallintaa molekyylejä, joita on tietty äärellinen määrä.

Laajennetun mH-systeemin (N, T, A, R) komponenteista N, T ja R ovat kuten laajennetuilla H-systeemeillä ja A on äärellinen *aksioomien* multijoukko yli joukon $A \subseteq V^+$, missä $V = N \cup T$. Multijoukkojen joukon $\mathcal{M}$ ehdoista

$$M_1 \Rightarrow_\gamma M_2 \quad \Leftrightarrow \quad \exists x, y, z, w \in (N \cup T)^+, r \in R :$$

(i) $M_1(x) \geq 1$, $(M_1 - \{(x, 1)\})(y) \geq 1$

(ii) $M_2 = ((M_1 - \{(x, 1)\}) - \{(y, 1)\}) \cup \{(z, 1)\} \cup \{(w, 1)\}$,

(iii) $(x, y) \vdash_r (z, w)$

määräytyvän relaation $\Rightarrow_\gamma$ refleksiivisen ja transitiivisen sulkeuman $\Rightarrow_\gamma^*$ avulla voidaan määritellä laajennetun mH-systeemin γ *generoima* kieli on

$$L(\gamma) = \{w \in T^* \mid \exists M' \in \mathcal{M},\ w \in Supp(M') :\ A \Rightarrow_\gamma^* M'\}. \tag{65}$$

Ehdon (i), joka ottaa huomioon tapauksen $x = y$, mukaan multijoukon M_1 sanoihin x ja y voidaan soveltaa silmukointisääntöä r vain, jos sanoja x ja y on vähintään yksi kappale. Ehdon (ii) mukaan tällöin sanojen x ja y lukumäärä vähenee yhdellä ja silmukoinnin tuloksena saatujen sanojen lukumäärä lisääntyy yhdellä.

Huomautus 3.7. Koska multijoukossa alkioita voi olla äärettömän monta, laajennettuja H-systeemejä voidaan pitää laajennettujen mH-systeemien erikoistapauksena, jossa jokaista aksioomaa on äärettömän monta.

Pykälän loppuosan tulokset ovat peräisin artikkelista [7].

Lause 3.8. *Jokaista 0-tyypin kielioppia G kohti on olemassa sellainen laajennettu mH-systeemi γ, että $L(\gamma) = L(G)$.*

Todistus. Olkoon $G = (N, T, S, P)$ 0-tyypin kielioppi, jonka johtosäännöt $u \to v$ on muunnettu lauseen 1.15 normaalimuotoon, jolloin ne toteuttavat ehdot

$$1 \leq |u| \leq 2, \quad 0 \leq |v| \leq 2.$$

Säännöt, joissa $u = v$, voidaan jättää pois, koska ne eivät vaikuta generoituun kieleen. Leimataan kieliopin G johtosäännöt ja kirjoitetaan $r : u \to v$, jos johtosäännön $u \to v$ leima on r. Merkitään $U = N \cup T$ ja käytetään leimojen joukosta merkintää L.

Olkoon $\gamma = (N', T, A, R)$ laajennettu mH-systeemi, jolla on välisymbolit

$$N' = N \cup \{X_1, X_2, Y, Z_1, Z_2\} \cup \{(r), [r] \mid r \in L\},$$

missä sanat $(r), [r]$ tulkitaan kirjaimiksi, aksioomat

$$A = \{(w_0, 1), (w_r, \infty), (w_\alpha, \infty), (w'_\alpha, \infty), (w_t, \infty) \mid \forall r : u \to v \in P,\ \forall \alpha \in U\},$$

$$\begin{aligned} w_0 &= X_1{}^2 Y S X_2{}^2 \\ w_r &= (r)v[r] \\ w_\alpha &= Z_1 \alpha Y Z_2 \\ w'_\alpha &= Z_1 Y \alpha Z_2 \\ w_t &= YY \end{aligned}$$

ja silmukointisäännöt

1. $\dfrac{\delta_1\delta_2 Y u \mid \beta_1\beta_2}{(r)v \mid [r]}$ $\quad \forall r : u \to v \in P, \beta_1, \beta_2 \in U \cup \{X_2\}, \delta_1, \delta_2 \in U \cup \{X_1\}$
2. $\dfrac{Y \mid u[r]}{(r) \mid v\alpha}$ $\quad \forall r : u \to v \in P, \alpha \in U \cup \{X_2\}$
3. $\dfrac{\delta_1\delta_2 Y \alpha \mid \beta_1\beta_2}{Z_1\alpha Y \mid Z_2}$ $\quad \forall \alpha \in U, \beta_1, \beta_2 \in U \cup \{X_2\}, \delta_1, \delta_2 \in U \cup \{X_1\}$
4. $\dfrac{\delta \mid Y\alpha Z_2}{Z_1 \mid \alpha Y \beta}$ $\quad \forall \alpha \in U, \delta \in U \cup \{X_1\}, \beta \in U \cup \{X_2\}$
5. $\dfrac{\delta\alpha Y \mid \beta_1\beta_2\beta_3}{Z_1 Y \alpha \mid Z_2}$ $\quad \forall \alpha \in U, \beta_1 \in U, \beta_2, \beta_3 \in U \cup \{X_2\}, \delta \in U \cup \{X_1\}$
6. $\dfrac{\delta \mid \alpha Y Z_2}{Z_1 \mid Y\alpha\beta}$ $\quad \forall \alpha \in U, \delta \in U \cup \{X_1\}, \beta \in U \cup \{X_2\}$
7. $\dfrac{\quad \mid YY}{X_1{}^2 Y \mid w}$ $\quad \forall w \in \{X_2{}^2\} \cup T\{X_2{}^2\} \cup T^2\{X_2\} \cup T^3$
8. $\dfrac{\quad \mid X_2{}^2}{Y^3 \mid \quad}.$

Tarkastellaan luokkiin 1-8 jaoteltujen silmukointisääntöjen toimintaa. Luokkien 1 ja 2 silmukointisääntöjen avulla voidaan simuloida kieliopin G johtoa sanan w_0 sisällä silloin, kun kirjain Y on sanassa w_0 sanan u vasemmalla puolella, missä $u \to v$ on kieliopin G johtosääntö. Erityisesti alussa, kun $w_0 = X_1^2 Y S X_2^2$, tämä ehto toteutuu. Olkoon w_0 muotoa

$$X_1^2 w_1 Y u w_2 X_2^2,$$

missä $r : u \to v$ on kieliopin G johtosääntö, ja oletetaan, että sanaa w_0 on yksi kappale. Soveltamalla luokan 1 silmukointisääntöä sanaan w_0 ja aksioomaan $(r)[v][r]$, jota on äärettömän monta, saadaan

$$(X_1^2 w_1 Y u | w_2 X_2^2, (r)v|[r]) \models_1 (X_1^2 w_1 Y u[r], (r)v w_2 X_2^2). \tag{66}$$

Silmukoimalla silmukoinnin 66 tuloksena syntyvät sanat keskenään käyttäen luokan 2 silmukointisääntöä saadaan

$$(X_1^2 w_1 Y | u[r], (r)|v w_2 X_2^2) \models_2 (X_1^2 w_1 Y v w_2 X_2^2, (r)u[r]). \tag{67}$$

Toinen tuloksena saaduista sanoista, $(r)u[r]$, ei voi osallistua myöhempiin silmukointeihin, koska muotoa $u \to u$ olevat johtosäännöt poistettiin kieliopista G alussa. Koska sanoja $X_1^2 w_1 Y u[r]$ ja $(r)v w_2 X_2^2$ syntyi silmukoinnissa 66 yksi kappale, ne häviävät silmukoinnissa 67. Näin ollen luokkien 1 ja 2 silmukointisäännöillä on korvattu sana

$$X_1^2 w_1 Y u w_2 X_2^2$$

sanalla

$$X_1^2 w_1 Y v w_2 X_2^2, \tag{68}$$

missä $u \to v$ on kieliopin G johtosääntö, ja sanaa 68, josta käytetään edelleen merkintää w_0, on jälleen yksi kappale kuten alkutilanteessa.

Luokkien 3 ja 4 silmukointisääntöjen avulla voidaan kirjainta Y siirtää sanassa w_0 oikealle. Silmukoimalla w_0 aksiooman $Z_1 \alpha Y Z_2$ kanssa, missä α on kirjaimen Y oikealla puolella oleva kirjain sanassa w_0, saadaan luokan 3 silmukointisääntöä käyttäen

$$(X_1^2 w_1 Y \alpha | w_2 X_2^2, Z_1 \alpha Y | Z_2) \models_3 (X_1^2 w_1 Y \alpha Z_2, Z_1 \alpha Y w_2 X_2^2). \tag{69}$$

Silmukoimalla silmukoinnin 69 tuloksena syntyneet sanat keskenään saadaan luokan 4 silmukointisääntöä käyttäen

$$(X_1^2 w_1 | Y \alpha Z_2, Z_1 | \alpha Y w_2 X_2^2) \models_4 (X_1^2 w_1 \alpha Y w_2 X_2^2, Z_1 Y \alpha Z_2). \tag{70}$$

Näiden kahden silmukoinnin tuloksena sana

$$X_1^2 w_1 Y \alpha w_2 X_2^2$$

korvautuu sanalla

$$X_1^2 w_1 \alpha Y w_2 X_2^2$$

eli kirjain Y on liikkuu kirjaimen verran oikealle. Silmukoinnin 70 tuloksena syntynyt sana $Z_1 Y \alpha Z_2$ yksi aksioomista, joita on ääretön määrä.

Luokkien 5 ja 6 silmukointisääntöjen avulla voidaan vastaavasti liikuttaa kirjainta Y sanassa w_0 vasemmalle käyttäen aksioomaa $Z_1 Y \alpha Z_2$. Silmukointien

$$\begin{aligned} (X_1^2 w_1 \alpha Y | w_2 X_2^2, Z_1 Y \alpha | Z_2) &\models_5 (X_1^2 w_1 \alpha Y Z_2, Z_1 Y \alpha w_2 X_2^2), && (71) \\ (X_1^2 w_1 | \alpha Y Z_2, Z_1 | Y \alpha w_2 X_2^2) &\models_6 (X_1^2 w_1 Y \alpha w_2 X_2^2, Z_1 \alpha Y Z_2) && (72) \end{aligned}$$

tuloksena sana

$$X_1^2 w_1 \alpha Y w_2 X_2^2$$

korvautuu sanalla

$$X_1^2 w_1 Y \alpha w_2 X_2^2 \tag{73}$$

eli kirjain Y liikkuu sanassa w_0 kirjaimen verran vasemmalle. Samalla operaation 71 tuloksena syntyneet sanat häviävät. Operaation 72 tuloksena syntyy sanan 73 lisäksi aksiooma, jota on systeemissä jo valmiiksi äärettömän monta.

Luokkien 7 ja 8 silmukointisääntöjen avulla voidaan poistaa kirjaimet X_1, X_2 ja Y sanasta

$$w_0 = X_1^2 w_1 Y w_2 X_2^2.$$

Tällöin tulee ensin siirtää kirjain Y sanan X_1^2 viereen silmukointisäännöillä 5 ja 6. Soveltamalla tämän jälkeen luokkien 7 ja 8 sääntöjä,

$$\begin{aligned} (X_1^2 Y | w_1 w_2 X_2^2, | YY) &\models_7 (X_1^2 YYY, w_1 w_2 X_2^2), \\ (X_1^2 YYY |, w_1 w_2 | X_2^2) &\models_8 (X_1^2 YYY X_2^2, w_1 w_2), \end{aligned}$$

saadaan tuotettua yksi kappale sanaa $w_1 w_2$.

Edellisten tarkastelujen nojalla laajennetulla mH-systeemilla γ voidaan silmukoida mikä tahansa kielen $L(G)$ sana eli $L(G) \subseteq L(\gamma)$. Näytetään vielä, että $L(\gamma) \subseteq L(G)$. Mekaanisella tarkistuksella voidaan todeta, että mitään aksioomista $w_r, w_\alpha, w'_\alpha, w_t$ ei voida silmukoida keskenään ja että edellä kuvatuissa silmukoinneissa, joissa sovelletaan luokkien 1, 3, 5 ja 7 silmukointisääntöjä, silmukoinnin tuloksena saatuihin sanoihin voidaan soveltaa pareittain ainoastaan luokkien 2, 4, 6 ja 8 silmukointisääntöjä. Esimerkiksi sanoihin, jotka on saatu soveltamalla luokan 1 silmukointisääntöjä, voidaan soveltaa ainoastaan luokan 2 silmukointisääntöä ja sama pätee pareille (3,4),(5,6) ja (7,8). Koska sanaa w_0 on yksi kappale, luokkien 1,3,5 ja 7 silmukointisääntöjä sovellettaessa saadut sanat häviävät luokkien 2,4,6 ja 8 silmukointisääntöjä sovellettaessa. Olkoon $w \in L(\gamma)$, jolloin $w \in T^*$. Ainoa tapa silmukoida sana yli pääteaakkoston on aloittaa sanasta w_0, mahdollisesti simuloida kieliopin G johtoa sanan w_0 sisällä ja poistaa välisymbolit X_1, X_2, Y sanasta w_0 käyttäen silmukointisääntöjä 7 ja 8. Tämän jälkeen kieliopin G simulointia ei voi jatkaa kirjaimen Y puuttumisesta johtuen, joten $w \in L(G)$. $\square$

Seuraavasta esimerkistä käy ilmi, että lauseen 3.8 mallin oikeellisuuden kannalta on olennaista, että aksioomaa w_0 on yksi kappale.

Esimerkki 3.9. Kieliopilla $G = (\{S\}, \{a, b\}, S, \{S \to aSa, S \to bSb, S \to \lambda\}$ voidaan johtaa tarkalleen parillispituiset palindromit yli aakkoston $\{a, b\}$. Olkoon $\gamma = (N', \{a, b\}, A, R)$ lauseen 3.8 todistuksen konstruktion mukainen kielioppia G vastaava laajennettu mH-systeemi, jolla on välisymbolit

$$N' = \{S, X_1, X_2, Y, Z_1, Z_2, (r_1), [r_1], (r_2), [r_2], (r_3), [r_3]\}$$

ja aksioomat

$$\begin{aligned} A \;=\; & \{(X_1^2 YSX_2^2, 1)\} \cup \\ & \{((r_1)aSa[r_1], \infty), ((r_2)bSb[r_2], \infty), ((r_3)[r_3], \infty)\} \cup \\ & \{(Z_1SYZ_2, \infty), (Z_1aYZ_2, \infty), (Z_1bYZ_2, \infty) \cup \\ & (Z_1YSZ_2, \infty), (Z_1YaZ_2, \infty), (Z_1YbZ_2, \infty) \cup \\ & (YY, \infty)\} \end{aligned}$$

ja taulukon 3.3 mukaiset silmukointisäännöt R.

Näytetään, miten laajennetulla mH-systeemillä γ voidaan simuloida kieliopin G johtoa

$$S \Rightarrow_G aSa \Rightarrow_G abSba \Rightarrow_G abba.$$

Silmukointien

$$\begin{aligned} (X_1^2YS|X_2^2, (r_1)aSa|[r_1]) &\models_1 (X_1^2YS[r_1], (r_1)aSaX_2^2) \\ (X_1^2Y|S[r_1], (r_1)|aSaX_2^2) &\models_2 (X_1^2YaSaX_2^2, (r_1)S[r_1]) \\ (X_1^2Ya|SaX_2^2, Z_1aY|Z_2) &\models_3 (X_1^2YaZ_2, Z_1aYSaX_2^2) \\ (X_1^2|YaZ_2, Z_1|aYSaX_2^2) &\models_4 (X_1^2aYSaX_2^2, Z_1YaZ_2) \\ (X_1^2aYS|aX_2^2, (r_2)bSb|[r_2]) &\models_1 (X_1^2aYS[r_2], (r_2)bSbaX_2^2) \\ (X_1^2aY|S[r_2], (r_2)|bSbaX_2^2) &\models_2 (X_1^2aYbSbaX_2^2, (r_2)S[r_2]) \end{aligned}$$

$$
\begin{aligned}
(X_1^2aYb|SbaX_2^2, Z_1bY|Z_2) &\models_3 (X_1^2aYbZ_2, Z_1bYSbaX_2^2)\\
(X_1^2a|YbZ_2, Z_1|bYSbaX_2^2) &\models_4 (X_1^2abYSbaX_2^2, Z_1YbZ_2)\\
(X_1^2abYS|baX_2^2, (r_3)|[r_3]) &\models_1 (X_1^2abYS[r_3], (r_3)baX_2^2)\\
(X_1^2abY|S[r_3], (r_3)|baX_2^2) &\models_2 (X_1^2abYbaX_2^2, (r_3)S[r_3])\\
(X_1^2abY|baX_2^2, Z_1Yb|Z_2) &\models_5 (X_1^2abYZ_2, Z_1YbbaX_2^2)\\
(X_1^2a|bYZ_2, Z_1|YbbaX_2^2) &\models_6 (X_1^2aYbbaX_2^2, Z_1bYZ_2)\\
(X_1^2aY|bba, Z_1Ya|Z_2) &\models_5 (X_1^2aYZ_2, Z_1YabbaX_2^2)\\
(X_1^2|aYZ_2, Z_1|YabbaX_2^2) &\models_6 (X_1^2YabbaX_2^2, Z_1aYZ_2)\\
(X_1^2Y|abbaX_2^2, |YY) &\models_7 (X_1^2Y^3, abbaX_2^2)\\
(X_1^2Y^3|, abba|X_2^2) &\models_8 (X_1^2Y^3X_2^2, abba)
\end{aligned}
$$

tuloksena saadaan yksi kappale sanaa *abba*.

Jos aksioomissa A olisi kaksi kappaletta sanaa $X_1^2YSX_2^2$, silmukointien

$$(X_1^2YS|X_2^2, (r_3)|[r_3]) \models_1 (\underline{X_1^2YS[r_3]}, (r_3)X_2^2)$$

ja

$$
\begin{aligned}
(X_1^2YS|X_2^2, (r_1)aSa|[r_1]) &\models_1 (X_1^2YS[r_1], (r_1)aSaX_2^2)\\
(X_1^2Y|S[r_1], (r_1)|aSaX_2^2) &\models_2 (X_1^2YaSaX_2^2, (r_1)S[r_1])\\
(X_1^2Ya|SaX_2^2, Z_1aY|Z_2) &\models_3 (X_1^2YaZ_2, Z_1aYSaX_2^2)\\
(X_1^2|YaZ_2, Z_1|aYSaX_2^2) &\models_4 (X_1^2aYSaX_2^2, Z_1YaZ_2)\\
(X_1^2aYS|aX_2^2, (r_3)|[r_3]) &\models_1 (X_1^2aYS[r_3], \underline{(r_3)aX_2^2})
\end{aligned}
$$

tuloksena saadaan yksi kappale ylläolevissa silmukoinneissa alleviivattuja sanoja $X_1YS[r_3]$ ja $(r_3)aX_2^2$. Silmukointien

$$
\begin{aligned}
(X_1^2Y|S[r_3], (r_3)|aX_2^2) &\models_2 (X_1^2YaX_2^2, (r_3)S[r_3])\\
(X_1^2Y|aX_2^2, |YY) &\models_7 (X_1^2Y^3, aX_2^2)\\
(X_1^2Y^3|, a|X_2^2) &\models_8 (X_1^2Y^3X_2^2, a)
\end{aligned}
$$

tuloksena syntyy yksi kappale sanaa $a \notin L(G)$.

Lemma 3.10. *Jokaista laajennettua mH-systeemiä γ kohti on olemassa korkeintaan kahdella aksioomalla varustettu laajennettu mH-systeemi γ', joka toteuttaa $L(\gamma') = L(\gamma)$.*

Todistus. Olkoon $\gamma = (N, T, A, R)$, laajennettu mH-systeemi, jonka aksioomien multijoukko A on äärellinen. Olkoot $w_1, w_2, \ldots, w_n$ ne joukon $supp(A)$

sanat, joita on äärellinen määrä ja $z_1, z_2, \ldots, z_m$ ne joukon $supp(A)$ sanat, joita on ääretön määrä. Muodostetaan laajennettu mH-systeemi

$$\gamma' = (N \cup \{c, d_1, d_2\}, T, A', R'),$$

missä kirjaimet c, d_1, d_2 ovat uusia välisymboleja, jolla on aksioomat

$$A' = \{((w_1c)^{A(w_1)}(w_2c)^{A(w_2)} \cdots (w_nc)^{A(w_n)}, 1), (d_1cz_1cz_2c \cdots cz_mcd_2, \infty)\}$$

ja silmukointisäännöt

$$R' = R \cup \{\#c\$d_2\#, \#d_1\$c\#\}.$$

Olkoon joukon A' ensimmäinen aksiooma w ja toinen z. Jos $n = 0$ eli multijoukossa A ei ole sanoja, joita on äärellisen monta, niin

$$w \notin supp(A')$$

ja jos $m = 0$ eli multijoukossa A ei ole sanoja, joita äärettömän monta, niin

$$z = d_1cd_2.$$

Näytetään, että laajennetulla mH-systeemillä γ' voi tuottaa äärettömän määrän kopioita sanoista z_j, $1 \leq j \leq m$. Silmukointien

$$\begin{aligned} & (d_1cz_1c \cdots cz_{j-1}c|z_jc \cdots cz_md_2, |z) \\ \models_{\#d_1\$c\#} \quad & (d_1cz_1c \cdots cz_{j-1}cz, z_jc \cdots cz_md_2), \\ & (z|, z_j|cz_{j+1}c \cdots cz_md_2) \\ \models_{\#c\$d_2\#} \quad & (zcz_{j+1}c \cdots cz_md_2, \underline{z_j}) \end{aligned}$$

tuloksena syntyy yksi kappale sanaa z_j. Prosessia voidaan toistaa mielivaltaisen monta kertaa, koska $A'(z) = \infty$.

Näytetään, että systeemillä γ' voidaan tuottaa tarkalleen $A(w_1)$ kappaletta sanaa w_1. Silmukointien

$$\begin{aligned} & ((w_1|c)(w_1c)^{A(w_1)-1}(w_2c)^{A(w_2)} \cdots (w_nc)^{A(w_n)}, z|) \\ \models_{\#c\$d_2\#} \quad & (\underline{w_1}, zc(w_1c)^{A(w_1)-1}(w_2c)^{A(w_2)} \cdots (w_nc)^{A(w_n)}), \\ & (zc|(w_1c)^{A(w_1)-1}(w_2c)^{A(w_2)} \cdots (w_nc)^{A(w_n)}, |z) \\ \models_{\#d_1\$c\#} \quad & (zcz, (w_1c)^{A(w_1)-1}(w_2c)^{A(w_2)} \cdots (w_nc)^{A(w_n)}) \end{aligned} \tag{74}$$

tuloksena syntyy yksi kappale sanaa w_1 ja prosessin aikana sana

$$(w_1c)^{A(w_1)} \cdots (w_nc)^{A(w_n)}$$

muuttuu sanaksi

$$(w_1c)^{A(w_1)-1}(w_2c)^{A(w_2)}\cdots(w_nc)^{A(w_n)},$$

jota on jälleen yksi kappale. Silmukoinnin 74 tuloksena syntynyt sana zcz sisältää vain aksioomia, joita voidaan tuottaa ääretön määrä. Vastaava prosessi voidaan suorittaa sanoille w_i, $2 \leq i \leq n$, joten laajennetulla mH-systeemillä γ' voidaan tuottaa tarkalleen $A(w_i)$ kappaletta sanoja w_i, $1 \leq i \leq n$. Näin ollen $L(\gamma) \subseteq L(\gamma')$.

Jos joukon R silmukointisääntöjä sovelletaan laajennetussa mH-systeemissä γ', silmukointi tapahtuu tällöin kirjainten c välissä olevalle sanalle. Koska joukon $R'\backslash R$ silmukointisäännöt pitävät kirjainten c välissä olevat sanat ehjinä, laajennetulla mH-systeemillä γ' ei voida silmukoida kieleen $L(G)$ kuulumattomia sanoja. Näin ollen $L(\gamma') \subseteq L(\gamma)$. □

Lemma 3.11. *Jokaista laajennettua mH-systeemiä γ kohti on olemassa kahdella välisymbolilla varustettu laajennettu mH-systeemi γ', joka toteuttaa $L(\gamma') = L(\gamma)$.*

Todistus. Olkoot laajennetun mH-systeemin $\gamma = (N, T, A, R)$ välisymbolit $N = \{N_1, \ldots, N_n\}$. Määritellään koodausta varten morfismi $h : (N \cup T)^* \rightarrow (\{c_1, c_2\} \cup T)^*$,

$$h(x) \quad = \quad \begin{cases} c_1c_2^ic_1, & \text{jos} \quad x = N_i, \quad 1 \leq i \leq n \\ a & \text{jos} \quad a \in T, \end{cases}$$

missä kirjaimet c_1 ja c_2 ovat aakkoston $N \cup T$ ulkopuolisia kirjaimia. Laajennettu mH-systeemi

$$\gamma' = (\{c_1, c_2\}, T, A', R'),$$

missä

$$A' = h(A)$$

ja

$$R' = \{h(u_1)\#h(u_2)\$h(u_3)\#h(u_4) \mid u_1\#u_2\$u_3\#u_4 \in R\},$$

toteuttaa $L(\gamma') = L(\gamma)$, koska johtuen siitä, että sanat $c_1c_2^ic_1$ pysyvät ehjinä silmukoinnin aikana, laajennettu mH-systeemi γ' toimii vastaavasti kuin γ. □

Laajennettuihin mH-systeemeihin voidaan liittää lisäinformaatiota aksioomiin lisätyssä sanassa. Jos $\gamma = (N, T, A, R)$ on laajennettu mH-systeemi ja $w \in (N \cup T)^*$, niin otetaan käyttöön merkintä

$$L(\gamma, w) = L((N, T, A \cup \{(w, 1)\})).$$

Lause 3.12. *On olemassa pääteaakkostolla T varustettujen laajennettujen mH-systeemien suhteen universaali laajennettu mH-systeemi.*

Todistus. Olkoon $\gamma_0 = (N, T, A, R)$ laajennettu mH-systeemi. Kielen $L(\gamma_0)$ sanat voidaan luetella tehokkaasti käymällä aksioomista lähtien läpi kaikki sanat, jotka laajennetulla mH-systeemillä γ_0 voidaan silmukoida järjestettynä sen suhteen, kuin monella silmukoinnilla kyseiset sanat on saatu, ja luettelemalla ne kieleen $L(\gamma_0)$ kuuluvaksi, jos kysymyksessä on sana yli aakkoston T. Näin ollen $L(\gamma_0) \in RE$.

Lauseen 1.25 nojalla on olemassa sellainen 0-tyypin kielioppi G_0, että $L(G_0) = L(\gamma_0)$. Lauseen 1.26 nojalla on olemassa pääteaakkostolla T varustettujen 0-tyypin kielioppien suhteen universaali 0-tyypin kielioppi

$$G_U = (N_U, T, S, P)$$

ja sellainen kieliopin G_U alkusymbolin S sisältävä sana $w(G_0) \in (N_U \cup T)^*$, että

$$L(G_U, w(G_0)) = L(G_0).$$

Olkoon $\gamma_1 = (N_1, T, A_1, R_1)$ laajennettu mH-systeemi, joka saadaan lauseen 3.8 todistuksen kieliopista G_U antamasta laajennetusta mH-systeemistä poistamalla sen aksioomista sana w_0. Tällöin aksioomissä A_1 ei ole yhtään sanaa, jota on äärellisen monta, joten soveltamalla lemman 3.10 todistusta saadaan laajennetusta mH-systeemistä γ_1 sen kanssa ekvivalentti laajennettu mH-systeemi

$$\gamma_2 = (N_2, T, A_2, R_2),$$

jolla on yksi aksiooma. Olkoon

$$\gamma_U = (\{c_1, c_2\}, T, A_U, R_U)$$

laajennettu mH-systeemi, joka on saatu laajennetusta mH-systeemistä γ_2 soveltamalla siihen lemman 3.11 todistusta.

Näytetään, γ_U on etsitty universaali laajennettu mH-systeemi. Koodataan sana $X_1^2 Y w(G_0) X_2^2$ sanaksi $w(\gamma_0)$ yli aakkoston $\{c_1, c_2\} \cup T$ käyttäen samaa morfismia, jolla laajennettu mH-systeemi γ_2 on koodattu laajennetuksi mH-systeemiksi γ_U. Tällöin

$$\begin{aligned} L(\gamma_U, w(\gamma_0)) &= L(G_U, w(G_0)) \\ &= L(G_0) \\ &= L(\gamma_0). \end{aligned}$$

□

Tässä pykälässä on näytetty, että yhden sanan lukumäärää rajoittamalla laajennettujen H-systeemien laskentavoima voidaan nostaa pykälässä 3.2 todistetusta äärellisten automaattien tasosta Turingin koneiden tasolle. Mitä tulee konstruoidun universaalin H-systeemin realistisuuteen, niin sen silmukointisääntöjen pituudet ovat samaa luokkaa kuin restriktioentsyymien pituudet, joista on esimerkkejä pykälän 3.1 taulukossa. Sen sijaan seuraavat universaalin mallin puutteet tekevät nopean rinnakkaislaskennan käytännössä mahdottomaksi.

(i) Mallin oikeellisuuden kannalta on olennaista, että sanaa w_0 on yksi kappale, mutta molekyylin reagointi on epätodennäköistä, jos sitä on vain yksi kappale.

(ii) Mallilla on paljon silmukointisääntöjä, mutta useiden restriktioentsyymien samanaikainen toiminta on mahdotonta, koska eri restriktioentsyymit saattavat vaatia eri reaktio-olosuhteet.

Viitteet

[1] L.M. Adleman: Molecular computation of solutions to combinatorial problems, *Science*, 226, 1021-1024, 1994

[2] J. Berstel: Transductions and context-free languages, B.G.Teubner, Stuttgart, 1979

[3] CSC tieteellinen laskenta OY: Bioinformatiikan sanasto, http://www.csc.fi/molbio/sanasto/html/index.html

[4] K. Culik, T. Harju: Splicing semigroups of dominoes and DNA, *Discrete Applied Mathematics*, 31, 261-277, 1991

[5] S. Eilenberg: Automata, languages and machines A, Academic Press, New York, 1974.

[6] R.P. Feynman: There's plenty of room at the bottom, http://www.zyvex.com/nanotech/feynman.html, 1960

[7] R. Freund, L. Kari, G. Pâun: DNA computing based on splicing: the existence of universal computers, *Theory of computing systems*, 32, 69-112, 1999

[8] T. Head: Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors, *Bulletin of Mathematical Biology*, 49, 737-759, 1987

[9] J. Karhumäki: Automata and formal languages, luentomoniste, 2005

[10] A. Mateescu, A. Salomaa: Aspects of classical language theory. G. Rozenberg, A. Salomaa (toim.): Handbook of formal languages, volume 1, Springer-Verlag, Berlin, 1997

[11] MBI Fermentas: Molecular biology catalog & product application guide, 2000-2001

[12] G. Pâun, G. Rozenberg, A. Salomaa: DNA computing, Springer-Verlag, Berlin, 1998.

[13] D. Pixton: Regularity of splicing languages, *Discrete applied mathematics*, 69, 101-124, 1996

[14] E. Post: A variant of a recursively unsolvable problem, *Bulletin of the American Mathematical Society*, 52, 264-268, 1946

[15] A. Salomaa: Formal languages, Academic Press, New York, 1973

[16] A. Salomaa: Jewels of formal language theory, Computer Science Press, Rockville, 1981

[17] D.B. Searls: Formal language theory and biological macromolecules, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 47, 117-141, 1999

[18] D.B. Searls: The linguistics of DNA, *American Scientist*, 80, 579-591, 1992

[19] P. Tirri, J. Lehtonen, R. Lemmetyinen, S. Pihakaski, P. Portin: Biologinen sanakirja, Otava, Helsinki, 2001