ÖREBRO
UNIVERSITET
Institutionen för naturvetenskap och teknik

Medianer under trecykelavståndet på S_n

Toni Duras och Niklas Lindman

Örebro universitet
Institutionen för naturvetenskap och teknik
Matematik C, 76 – 90 högskolepoäng

Medianer under trecykelavståndet på S_n

Toni Duras och Niklas Lindman

April 2012

Handledare: Niklas Eriksen
Examinator: Holger Schellwat

Självständigt arbete, 15 hp
Matematik, C–nivå, 76 – 90 hp

Sammanfattning

Att bestämma släktskap mellan olika arter är viktigt för att förstå hur dessa utvecklas med tiden. Små ändringar på nukleotidnivå kan innebära stora skillnader och egenskapsförändringar hos en art. Därför är forskning kring detta viktigt för att förstå olika arters släktskap. Med hjälp av blocktranspositioner går det att se avståndet mellan olika permutationer och detta görs genom att hitta medianer. En median är en genordning sådan att summan av blocktranspositionsavståndet till de tre genordningarna minimeras. En blocktransposition har tre brytpunkter, och på samma sätt kommer blocktranspositionsavståndet att ha egenskaper som liknar 3-cykelavståndet på permutationer. Därför kommer vi i denna uppsats att undersöka detta och väljer alltså att studera 3-cykelavståndet på permutationer och dess medianer.

Innehåll

1 Introduktion **5**

2 Bakgrund och definitioner **9**
2.1 Medianer och trecykelavstånd 9
2.2 Grafer 10
2.3 Beräkning av trecykelavstånd med hjälp av grafer 12

3 Stigar och cykler i reducerade grafer **15**

Kapitel 1

Introduktion

Att bestämma släktskap mellan olika arter kan vara svårt. Ett välkänt exempel är blåsippan, som Carl von Linné gav namnet *Anemone hepatica* och placerade i samma släkte som vitsippan (*Anemone nemorosa*). Senare forskning flyttade ut blåsippan och dess närmsta släktingar till ett angränsande släkte och bytte det latinska namnet till *Hepatica nobilis*. Modern DNA-forskning signalerar dock att de två familjerna borde slås ihop igen.

Att titta på DNA är ofta det säkraste sättet att fastslå släktskap mellan organismer, särskilt mellan till exempel bakterier som inte har så tydliga utseendemässiga egenskaper att gå efter. DNA består av en lång sträng av nukleotider som kan grupperas i funktionella enheter, gener. Likheter i uppbyggnaden av DNA antyder att två arter är nära släkt, olikheter att släktskapet är mer avlägset.

Skillnader mellan två arter beror huvudsakligen på skillnader på nukleotidnivå. En förändring av bara några få nukleotider kan göra att en gen slutar fungera eller ändrar egenskaper fullständigt. Studier av dessa förändringar är viktiga om man ska se hur arter förändras med tiden. Däremot kan det bli problem att bestämma släktskap mellan arter genom att följa utvecklingen hos en gen, eftersom gener kan hoppa mellan arter. Vi kan bestämma hur genen har utvecklats, men det behöver inte stämma överens med hur arterna utvecklats.

Ett annat alternativ är att studera hur ordningen mellan generna utvecklats. Detta har enbart en liten inverkan på arternas utveckling, men förändringar i genordning kan på samma sätt som förändringar i nukleotider användas för att avgöra släktskap. Små skillnader indikerar nära släktskap och stora skillnader indikerar avlägset släktskap. När vi tittar på bakterier är skillnaden i genuppsättning så liten att vi helt enkelt kan strunta i den och titta bara på de gener som förekommer exakt en gång i båda bakterierna vi jämför.

Studier av avståndsberäkningar mellan genordningar har varit ett forskningsområde i ungefär 20 år. En översikt över forskningsläget 2009 ges i

[3]. Genordningen modelleras som en permutation. Ordningen mellan generna kan ändras på några olika sätt. De viktigaste förändringarna på **genom** (arvsmassan hos en organism) med en **kromosom** (sträng av nukleotider) är vändningar och blocktranspositioner. En **vändning** innebär att ett segment av gener tas ut och placeras tillbaks på samma plats fast baklänges, till exempel 1<u>234</u>5 dsaokdjsaolkdjoaksjd som kan ändras till 1<u>432</u>5. En **blocktransposition** innebär att ett segment tas ut och placeras in på en annan plats, till exempel 12<u>34</u>5 som då kan ändras till 1<u>34</u>25 [1]. Blocktranspositionen kan också ske genom att det uttagna blocket placeras in baklänges på den nya platsen.

Avstånd mellan två olika genordningar beräknas i antalet vändningar eller blocktranspositioner som krävs för att omvandla den ena genordningen till den andra. Tidigare har huvudsakligen vändningar studerats, eftersom vändningsavstånd är enklare att beräkna än blocktranspositionsavstånd, men i denna uppsats har vi som avsikt att studera just blocktranspositioner. Mer om vändningar, blocktranspositioner och sätt att mäta evolutionära avstånd finner ni i [3].

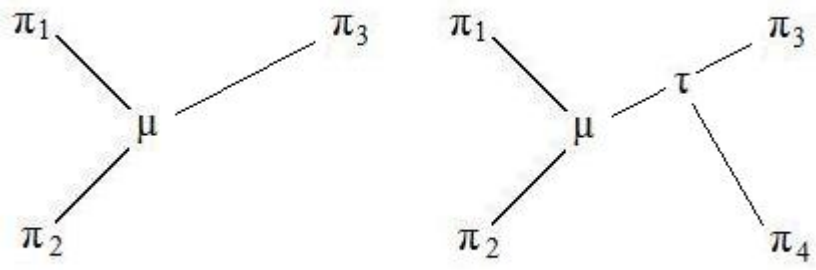
En **median** av k givna genordningar är en genordning sådan att summan av blocktranspositionsavståndet till de k genordningarna minimeras. Om mängden av de givna genordningarna betecknas $S = \{\pi_1, \pi_2, \ldots, \pi_k\}$ ges avståndet mellan en genordning π och mängden S av

$$d(\pi, S) = \sum_{i=1}^{k} d(\pi, \pi_i).$$

Genordningen μ är en median om μ minimerar $d(\mu, S)$.

Medianer kan användas för att bestämma släktträd för arter. Ett genetiskt släktträd kan alltså byggas upp av att undersöka medianer mellan tre olika genordningar. Givet ett antal arters genordningar tas först en median fram för tre godtyckliga genordningar. Sedan görs samma sak fast med en ny genordning och alla val av två av de tidigare. Resultatet i var och en av dessa beräkningar kan då bli samma median som tidigare eller en annan median. Det fall där vi får en annan median visar var den nya medianen ska sättas in. Om permutationerna π_1, π_2 och π_3 ger medianen μ och π_1, π_3 och π_4 ger medianen τ betyder det att τ bryter grenen mellan μ och π_3 (se figur 1.1).

Vi skrev ovan att det är svårt att beräkna blocktranspositionsavstånd och därmed borde det vara i stort sett omöjligt att beräkna blocktranspositionsmedianer. Det verkar det också vara, men vi kommer att titta på något som är väldigt likt dessa medianer, men som borde vara betydligt enklare att beräkna. Det visar sig nämligen att tillväxttakten hos vändningsavståndet mellan två genordningar som utsätts för vändningar är i stort sett samma som tillväxttakten hos transpositionsavståndet (en **transposition** innebär att två element byter plats med varandra) mellan två permutationer som utsätts för transpositioner [2]. Detta avstånd kallas också för 2-cykelavståndet,



Figur 1.1: Median av $\pi_1, \pi_2, \pi_3 = \mu$, median av $\pi_1, \pi_2, \pi_4 = \mu$, median av $\pi_1, \pi_3, \pi_4 = \tau$, median av $\pi_2, \pi_3, \pi_4 = \tau$.

eftersom en transposition på cykelform blir en 2-cykel.

De två elementen i 2-cykeln svarar mot de två **brytpunkterna** i en vändning, det vill säga de två ställen där genordningen tas isär och sätts ihop igen. En blocktransposition har tre brytpunkter, och på samma sätt kommer blocktranspositionsavståndet att ha egenskaper som liknar 3-cykelavståndet på permutationer. Det vill säga ha samma tillväxttakt. Vi väljer därför att studera 3-cykelavståndet på permutationer och dess medianer.

En median skapas enklast genom att tilldela värden till en position i taget i den tänkta medianen. Detta fortgår antingen tills vi har en färdig permutation eller tills vi ser att hur vi än fortsätter så kommer medianavståndet bli större än det bästa vi funnit hittills. På detta sätt gör vi en fullständig sökning genom samtliga permutationer. Syftet med arbetet är att dels finna en undre gräns för 3-cykelmedianavståndet mellan en mängd S av permutationer och ett godtyckligt färdigställande av en halvfärdig mediankandidat, dels att undersöka om det går att dra några generella slutsatser om vilka värden en median kan ha på sina positioner.

En skarp undre gräns är viktigt för att snabbt kunna avfärda mediankandidater som inte kommer att bli medianer oavsett hur vi tilldelar de resterande positionerna. Vi kommer att jobba med en graf där varje ny tilldelning svarar mot en svart kant. När vi drar ett svart kant kan vi se om den undre gränsen har förändrats. På så sätt kan vi bilda oss en uppfattning om vilka svarta kanter som ingår i medianen och vilka som inte ingår.

I avsnitt 2 går vi igenom definitioner och lemman som ligger till grund för detta arbete. De flesta definitioner vi använder oss av är hämtade från [1]. Mycket går att överföra utan större problem, men dessvärre har vi stött på patrull med nyckellemmat som ger övre och undre gränser för en halvfärdig median. Den naturliga generaliseringen visar sig inte stämma, vilket vi visar med ett motexempel.

Vi har ändå valt att fortsätta med analysen, med förhoppningen att andra undre och övre gränser kan dra nytta av den analys vi gör. Avsnitt 3 innehåller denna analys som är modellerad efter analysen av tvåcykelavståndet i [1]. Vi finner motsvarigheter till samtliga lemman i tvåcykelfallet. Vi tvingas dock konstatera att medan det för tvåcykelmedianen räcker att ett element befinner sig på en viss position i hälften av de genom som betraktas för att detta element ska ha denna position i alla medianer, så gäller för trecykelavstånd att elementet måste vara på samma position i samtliga genom för att garanterat vara med i åtminstone vissa medianer. Trecykelmedianer är således mycket svårare att beräkna.

Kapitel 2

Bakgrund och definitioner

2.1 Medianer och trecykelavstånd

Låt $S = \{\pi_1, \pi_2, \ldots, \pi_k\}$, $\pi_i \in S_n$, vara en mängd av permutationer kallad **basmängd**. Vi kommer att använda oss av linjär notation (exempelvis $\pi = 3412$) och cykelnotation ($\pi = (1\ 3) \circ (2\ 4)$). Om inte annat anges är k antalet permutationer i S och n längden av permutationerna. En avståndsfunktion mellan två permutationer noteras $d(\cdot, \cdot)$ och avståndet mellan en permutation $\pi \in S_n$ och S defineras som

$$d(\pi, S) = \sum_{i=1}^{k} d(\pi, \pi_i).$$

En **median** är något $\mu \in S_n$ som minimerar avståndet $d(\mu, S)$. Mängden av medianer noteras $M(S)$ och vi låter $d(S) = d(\mu, S)$ för $\mu \in M(S)$. Valet av avståndsmätning $d(\cdot, \cdot)$ ger upphov till flera intressanta medianproblem; i denna artikel kommer vi fokusera på medianproblemet för **trecykelavståndet**, förkortat d_3.

Det är välkänt att följande begränsningar för $d(S)$ gäller för metriska avstånd [1].

Lemma 2.1.1. *För varje avståndsmått $d(\cdot, \cdot)$ är medianavståndet $d(S)$ för $S = \{\pi_1, \ldots, \pi_k\}$ begränsat av*

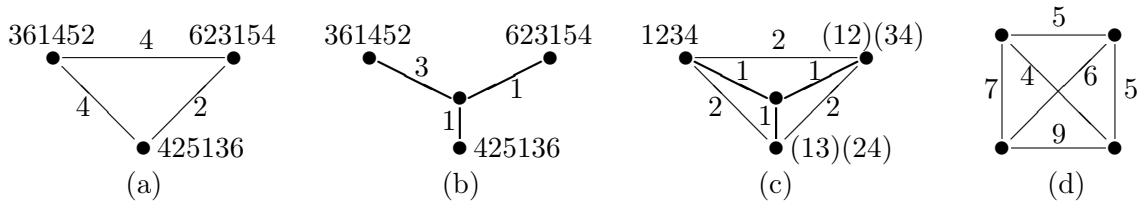
$$\frac{\sum_{i<j} d(\pi_i, \pi_j)}{k-1} \leq d(S) \leq \min_i \sum_j d(\pi_i, \pi_j).$$

Bevis. För den nedre gränsen ger triangelolikheten att $d(\pi_i, \pi_j) \leq d(\mu, \pi_i) + d(\mu, \pi_j)$, och därmed är $\sum_{i<j} d(\pi_i, \pi_j) \leq (k-1) \sum_i d(\mu, \pi_i)$. Den övre gränsen är minimum av $d(\pi_i, S)$ [1]. □

Noterbart är att den övre gränsen ger en uppskattning av $d(S)$ med ett fel som begränsas uppåt av faktorn $(2 - 2/k)$. Medianproblemet är således

trivialt för $k \leq 2$. Låt $D = \sum_{i=1}^{k} \sum_{j=1}^{k} d(\pi_i, \pi_j)$, det vill säga summan av avstånden mellan varje ordnat par av genom π_i och π_j. Låt dessutom L och U vara lägre och övre gränsen för $d(S)$. Då fås att $D/(k-1) = 2L$. Den övre gränsen U kan aldrig kan vara större än medelvärdet för alla permutationsavstånd, likhet fås om och endast om alla avstånd är lika stora. Det leder till att $kU \leq D$. Ur detta fås $kU \leq D = 2L(k-1) = L(2k-2) \iff U \leq \frac{L(2k-2)}{k} = L(2-2/k)$.

Exempel 2.1.1. Betrakta de tre permutationerna i (a) med givna transpositionsavstånd. Trädets undre gräns framgår i (b) där medianen med den undre gränsen är $\mu = 423156$. Basmängden S i (c) har dock $d_{\text{trp}}(S)$ som är strikt större än den undre gränsen. Detta innebär att $M(S) = S$ vilket i sin tur medför att $d_{\text{trp}}(S) = 4$, medan den undre gränsen är 3. Med de givna avstånden i (d) är den undre gränsen 12 vilket inte kan erhållas eftersom de övre och undre kanterna kräver att $d_{\text{trp}}(S) \leq 14$.

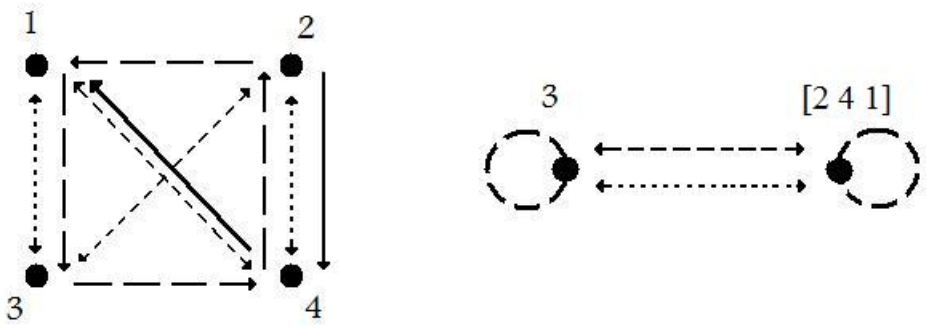


Betrakta bild (a). Skriver vi 361452 på cykelform får vi $(1\ 3) \circ (2\ 6) \circ (4) \circ (5)$ och skriver vi om 623154 fås $(1\ 6\ 4) \circ (2) \circ (3) \circ (5)$. Med hjälp av fyra transpositioner kan vi komma från den ena permutationen till den andra (t.ex. $(1\ 3) \circ (2\ 6) \circ (4) \circ (5) \circ (3\ 1) \circ (6\ 2) \circ (1\ 6) \circ (6\ 4) = (1\ 6\ 4) \circ (2) \circ (3) \circ (5)$), således är $d_2 = 4$. Vi kommer se senare att der inte går med färre transpositioner.

2.2 Grafer

En **graf** är en geometrisk tolkning av permutationer där grafens noder är element i permutationerna. Varje transposition i permutationerna blir en kant i grafen. I det följande hänvisar termen graf till grafer med färgade kanter $G = (V(G), E(G))$, om ingenting annat sägs, där $V(G)$ är en mängd av noder (även kallad vertex) och $E(G)$ är en mängd av kanter. En kant från v_1 till v_2 med färgen i noteras $(v_1 \xrightarrow{i} v_2)$. Antalet kanter från u till v noteras $|(u \longrightarrow v)|_G$. En **stig** är en serie noder sådana att $(v_1 \longrightarrow v_2 \longrightarrow \ldots \longrightarrow v_m)$, där $i \neq j \Rightarrow v_i \neq v_j$ och en **cykel** är en stig med $v_m = v_1$. **Längden** hos såväl stigar som cykler mäts i antalet noder, och en **udda stig** och en **udda cykel** har båda udda längd.

En **alternerande stig** med färgerna c_1 och c_2 i en graf G är en serie noder $v_1, v_2, \ldots, v_{2m}$ så att G innehåller kanter $(v_{2i-1} \xrightarrow{c_1} v_{2i})$ med färgen c_1 för $1 \leq i \leq m$ och kanter $(v_{2i} \xleftarrow{c_2} v_{2i+1})$ med färgen c_2. **Längden** m



Figur 2.1: En cykelgraf $G(S, \mu_2)$ med $S = \{\pi_1 = (1\ 3\ 4\ 2), \pi_2 = (1\ 3) \circ (2\ 4), \pi_3 = (1\ 4) \circ (2\ 3)\}$ och $\mu_2 = 4 \cdot\cdot 2$, där en punkt på position i indikerar att $i \notin \mu_2$, och dess reducerade graf red(G).

hos alternerande stigar och cykler ges av hälften av antalet noder, och som tidigare definieras **udda alternerade stigar och cykler** av att deras längd är udda. En **maximal alternerande stig** är en alternerande stig som inte kan förlängas.

Låt μ_b vara en partiell permutation med b svarta streck. När μ_b slutförs till μ_n fås en median till den givna permutationen. b är ett index på μ som talar om antal svarta streck.

Cykelgrafen av S, $G = G(S, \mu_b)$, är en graf med n noder märkta $1, \ldots, n$, med $(v_1 \xrightarrow{i} v_2)$ om $\pi_i(v_1) = v_2$. Det motsvarar brytpunktsgrafen som ofta betraktas när man studerar omvända avståndsproblem, men som har riktade kanter istället för oriktade. Cykelgrafen innehåller också b kanter med färgen $k + 1$, från och med nu kallar vi dom svarta, som indikerar inversen av den partiella permutationen μ_b: vi har $(v_1 \xrightarrow{k+1} v_2)$ om $\mu_b(v_2) = v_1$.

Om en cykelgraf $G = G(S, \mu_b)$ med b svarta kanter är given är den **reducerade cykelgrafen** $G' = \text{red}(G)$ definerad på följande sätt. För varje maximal stig $(v_1 \xrightarrow{k+1} v_2 \xrightarrow{k+1} \cdots \xrightarrow{k+1} v_m)$ i G, får vi noden $[v_1\ v_2 \ldots v_m]$ i G'. För varje maximal alternerande stig $(v_1 \xrightarrow{i} v_2 \xrightarrow{k+1} \cdots \xrightarrow{i} v_{2m})$ i G, lägger vi till kanten $[v_1 \ldots] \xrightarrow{i} [\ldots v_{2m}]$ i G'. Notera att eftersom den alternerande stigen är maximal, har vi ingen svart kant till v_1. Vi noterar att den reducerade cykelgrafen $G' = \text{red}(G)$ är en cykelgraf med $n - b$ noder, se [1].

Exempel 2.2.1. Betrakta cykelgrafen G till vänster i Figur 2.1, med $k = 3$ och två svarta kanter. Med svarta kanterna $(2 \xrightarrow{4} 4 \xrightarrow{4} 1)$ får vi noden

[241] i red(G) till höger i figuren. Med de långa strecken som färg 1, får vi den maximala alternerande stigen ($3 \xrightarrow{1} 4 \xleftarrow{4} 1 \xrightarrow{1} 3$) och ($2 \xrightarrow{1} 1$) in G, vilket ger kanterna ($3 \xrightarrow{1} 3$) och ($[4\ 2\ 1] \xrightarrow{1} [2\ 4\ 1]$) i red($G$).

Vi har att $c(\pi_i^{-1}\pi_j) = c(G, i, j)$, där $c(G, i, j)$ är antalet alternerande cykler i grafen G med färgerna i och j, antalet alternerande cyklar med färgerna svart $k+1$ och en färg i noteras $c(G, i)$. Antalet udda alternerande cykler noteras $c_{\mathrm{u}}(G, i, j)$ [1]. Vi låter $p(G, i)$ vara antalet maximala alternerande stigar och cykler med färgerna $k+1$ och i i G, och $p_{\mathrm{u}}(G, i)$ låter vi vara antalet udda maximala alternerande stigar och udda cykler med färgerna $k+1$ och i i G.

Lemma 2.2.1. *När ett svart streck dras mellan två stigar i G kommer stigarna att slås ihop till en stig i G'. För detta senario finns tre fall. Två jämna stigar slås ihop till en jämn, två udda stigar slås ihop till en jämn eller en udda och en jämn stig slås ihop till en udda stig.*

Bevis. Låt v_1 och v_2 ingå i två disjunkta stigar och $\mu_{b+1}(v_2) = v_1$ i G så kommer stigarna att slås ihop och bilda en ny stig G'. Det vill säga ($\ldots \longrightarrow v_1 \longrightarrow \ldots$) och ($\ldots \longrightarrow v_2 \longrightarrow \ldots$) i G slås ihop till ($\ldots \longrightarrow [v_1\ v_2] \longrightarrow \ldots$) i G'. Längden av den nya stigen blir då summan av längden av stigen till v_1 och längden av stigen från v_2. □

2.3 Beräkning av trecykelavstånd med hjälp av grafer

Att beräkna transpositionsavståndet eller 2-cykelavståndet är inte särskilt svårt. Följande sats är känd sedan länge [1].

Sats 2.3.1. *Transpositionsavståndet mellan $\sigma \in S_n$ och $\tau \in S_n$ fås av*

$$d_2(\sigma, \tau) = n - c(\sigma^{-1}\tau),$$

där $c(\pi)$ är antalet cykler i π.

Vi går nu vidare till 3-cykelavståndet, som inte studerats lika mycket tidigare. Detta avstånd går att beräkna på liknande sätt. För att kunna beräkna 3-cykelavståndet för udda permutationer, som till skillnad från de jämna permutationerna inte kan skrivas som en produkt av 3-cykler, definierar vi **3-cykelavståndet** $d_3(\sigma, \tau)$ mellan $\sigma \in S_n$ och $\tau \in S_n$ som det minsta antalet två- och trecykler som multiplicerade med σ ger τ.

Sats 2.3.2. *Trecykelavståndet, d_3, mellan $\sigma \in S_n$ och $\tau \in S_n$ fås av*

$$d_3(\sigma, \tau) = \frac{n - c_{\mathrm{u}}(\sigma^{-1}\tau)}{2},$$

där $c_{\mathrm{u}}(\pi)$ är antalet udda cykler i π.

Bevis. Man inser att om a, b och c tillhör samma cykel kommer antingen $(a\ b\ c)$ eller $(a\ c\ b)$ dela cykeln i tre delar. Det är det snabbaste sättet att öka antalet cykler. Antalet cykler kan alltså inte öka med mer än 2 när vi multiplicerar med en 3-cykel. Därmed minskar det givna avståndsformeln inte med mer än 1 när vi multiplicerar med en 3-cykel.

Vi vill visa att varje faktorisering av $\sigma^{-1}\tau$ i enbart transpositioner/3-cykler har $\frac{n-c_{\mathrm{u}}(\sigma^{-1}\tau)}{2}$ faktorer. Det gör vi genom att dela in i fall, så att något fall alltid gäller.

1. Om vi har en udda cykel av längd 3 eller längre kan den transformeras till tre udda cykler.

2. Om vi har en jämn cykel av längd större än 2 kan den transformeras till en jämn och två udda cykler.

3. Två 2-cykler kan transformeras till två udda cykler.

4. Har vi bara en 2-cykel kan den inte transformeras till någonting användbart av en trecykel. Däremot delas den av en tvåcykel.

Fall 1–2 illustreras av

$$(a_1 \dots a_{i-1}\ a_i \dots a_{j-1}\ a_j \dots a_m) \circ (a_1\ a_j\ a_i) = (a_1\ a_{j+1} \dots a_m) \circ (a_2 \dots a_i) \circ (a_{i+1} \dots a_j), 1 \leq i < j \leq m.$$

och fall 3 av

$$(a\ b) \circ (c\ d) \circ (a\ c\ b) = (a\ d\ c) \circ (b).$$

I samtliga fall ökar antalet udda cykler med 2. Därmed ger formeln avståndet. □

Exempel 2.3.1. Exempel på de olika fallen i Sats 2.3.2.
Fall 1: Cykeln (1 3 2 5 4) kan till exempel transformeras med hjälp av (1 2 3). $(1\ 3\ 2\ 5\ 4) \circ (1\ 2\ 3) = (1\ 5\ 4) \circ (2) \circ (3)$.
Fall 2: Cykeln (1 4 2 3) kan till exempel transformeras med hjälp av (1 2 4). $(1\ 4\ 2\ 3) \circ (1\ 2\ 4) = (1\ 3) \circ (2) \circ (4)$.
Fall 3: Cyklerna $(1\ 2) \circ (3\ 4)$ kan till exempel transformeras med hjälp av (1 3 2). $(1\ 2) \circ (3\ 4) \circ (1\ 3\ 2) = (1\ 4\ 3) \circ (2)$.

Lemma 2.3.1. *För en cykelgraf $G = G(S, \mu_b)$, gäller att d_3 mellan ett slutfört μ av μ_b och $\pi_i \in S$ uppfyller*

$$d_3(\mu, \pi_i) \geq \frac{n - p_{\mathrm{u}}(G, i)}{2}.$$

Bevis. För $b = 0$ är $p_{\mathrm{u}}(G(S, \mu_0), i) = n$ vilket ger $d_3 \geq 0$ och för $b = n$ är $p_{\mathrm{u}}(G(S, \mu_n), i) = c_{\mathrm{u}}(G(S, \mu_n), i)$. Då $b \neq \{0, n\}$ gäller det att $p_{\mathrm{u}}(G(S, \mu_{b-1}), i) - p_{\mathrm{u}}(G(S, \mu_b), i) = \{0, 2\}$ därför att när ett svart streck dras sluts en stig till en cykel eller så binds två stigar ihop. Alltså kan inte antalet udda stigar/cykler öka. □

De förberedelser som gjorts ovan syftar till att finna en motsvarighet en motsvarighet till Lemma 3.3 i [1]:

Lemma 2.3.2. *Låt $S = \{\pi_1, \pi_2, \ldots, \pi_k\}$ och låt $G = G(S, \mu_b)$ och $G' = \mathrm{red}(G)$. För godtyckligt slutfört μ av μ_b, har vi att*

$$\frac{\sum_{i<j}((n-b) - c(G', i, j))}{k-1} \leq d_2(\mu, S) - \sum_i (n - p(G, i)) \leq \min_i \sum_j ((n-b) - c(G', i, j)).$$

Denna formel ger oss gränser för avståndet till S hos de mediankandidater μ som man kan få från μ_b genom att räkna cykler och stigar i den reducerade grafen. På så sätt kan vi använda samma algoritm för att finna nästa kant genom hela beräkningen av μ, eftersom vi alltid tittar på den reducerade grafen utan svarta kanter. Motsvarande formel för 3-cykelavståndet skulle bli

$$\frac{\sum_{i<j} \frac{((n-b) - c_{\mathrm{u}}(G', i, j))}{2}}{k-1} \leq d_3(\mu, S) - \sum_i \frac{(n - p_{\mathrm{u}}(G, i))}{2} \leq \min_i \sum_j \frac{(n-b) - c_{\mathrm{u}}(G', i, j)}{2}$$

Dessvärre gäller inte denna formel. Ett exempel på det är fallet

$$\begin{aligned} \pi_1 &= (1\ 2\ 5) \circ (3\ 6\ 4) \\ \pi_2 &= (1\ 2\ 3\ 4\ 5\ 6) \\ \pi_3 &= \mathrm{id}. \end{aligned}$$

Ekvationen ger då $7/2 \leq d_3(\mu, S) \leq 4$ för alla μ, så medianavståndet $d(S)$ är 4. Ett enkelt sätt att uppfylla detta är att välja π_1 som median. Lägger vi nu in kanten $\mu(2) = 5$, vilket fortfarande kan leda till medianen π_1, får vi

$$\begin{aligned} \pi_1 &= (1\ [2\ 5]) \circ (3\ 6\ 4) \\ \pi_2 &= (1\ [2\ 5]\ 6) \circ (3\ 4) \\ \pi_3 &= \mathrm{id}. \end{aligned}$$

Ekvationen ger nu $5/2 \leq d_3(\mu, S) - 2 \leq 3$, vilket ger $d(\mu, S) = 5$. Fortsätter vi att lägga in fler kanter som leder till medianen π_1 återgår avståndet till 4.

Vi har inte funnit någon annan variant som kan ersätta detta lemma. Trots detta har vi fortsatt att finna motsvarigheter till många av de lemman och satser som ges i [1], som anger hur antalet cykler och stigar förändras då vi lägger till en kant. Vi beräknar vad som händer med antalet udda cykler och udda stigar, med förhoppningen att de kan komma till användning så småningom.

Kapitel 3

Stigar och cykler i reducerade grafer

Definitionerna och lemmana från föregående avsnitt är grundläggande för att i detta avsnitt kunna behandla arbetets egentliga syfte, nämligen att om möjligt finna en undre gräns för d_3 och dra eventuella slutsatser om vilka transpositioner som ingår i en median.

Lemma 3.0.3. *Antag att vi sätter $\mu_{b+1}(v_2) = v_1$, det vill säga att vi sätter en svart kant från v_1 till v_2 i $G(S, \mu_b)$ och får $G(S, \mu_{b+1})$. Då är*

$$\sum_i p_{\mathrm{u}}(G(S, \mu_b), i) - p_{\mathrm{u}}(G(S, \mu_{b+1}), i)) = 2 \sum_i a_i,$$

där $a_i = 1$ om $\xrightarrow{i} (\ldots v_1)$ och $(v_2 \ldots) \xrightarrow{i}$ är två disjunkta udda stigar och $a_i = 0$ annars.

Bevis. När ett svart streck dras finns 4 möjliga fall som kan påverka p-värdet av udda cykler och udda maximala alternerande stigar.

1. En maximal alternerande stig kan slutas till en cykel, då ändras inte p-värdet. Om stigen är udda blir cykeln udda och om stigen är jämn blir cyklen jämn.

2. Två jämna stigar kan bindas ihop till en stig som då också är jämn och således inte ändrar p-värdet.

3. En jämn och en udda stig kan bindas ihop till en udda stig vilket inte heller ändrar p-värdet.

4. Det intressanta fallet är om två udda stigar binds ihop eftersom de då bildar en jämn stig och därför minskar p-värdet med 2.

□

Lemma 3.0.4. *Antal udda maximala alternerande stigar och udda cykler ges av*

$$p_{\mathrm{u}}(G(S,\mu_b),i) = n - 2\sum_k p_k \left\lfloor \frac{k}{2} \right\rfloor$$

där vi har p_k stigar och cykler av längd k.

Bevis. I en graf med n stycken noder är n det maximala värdet på $p_{\mathrm{u}}(G(S,\mu_b),i)$. För varje svart streck som binder ihop två stycken udda maximala stigar till en jämn minskar $p_{\mathrm{u}}(G(S,\mu_b),i)$ med 2. Övriga svarta streck påverkar inte värdet på $p_{\mathrm{u}}(G(S,\mu_b),i)$. Samtliga fall med alternerande stigar av längd större än 3 kan ses som 2-stigar som sitter ihop om det är en jämn alternerande stig och 2-stigar som sitter ihop plus en 1-stig om det är en udda alternerande stig. □

Lemma 3.0.5. *Om kanterna* $([v_2\ldots] \xrightarrow{c_1} [\ldots v_1])$ *och* $([v_2\ldots] \xrightarrow{c_2} [\ldots v_1])$ *tillhör* $E(\mathrm{red}(G(S,\mu_b)))$*, då ger* $\mu_{b+1}(v_2) = v_1$ *att*

$$c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_b)),c_1,c_2) - c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_{b+1})),c_1,c_2) = \begin{cases} 0 & \textit{om cykeln är jämn} \\ 1 & \textit{om cykeln är udda} \end{cases}$$

Bevis. Den alternerande cykeln $([v_2\ldots] \xrightarrow{c_1} [\ldots v_1] \xleftarrow{c_2} [v_2\ldots])$ i $\mathrm{red}(G(S,\mu_b))$ kommer att försvinna och ingen annan cykel med färgerna c_1 och c_2 påverkas. Om cykeln är jämn sker ingen förändring i antalet udda cykler. Om cykeln är udda minskar antalet udda cykler med 1. □

Lemma 3.0.6. *Om* $([v_2\ldots] \xrightarrow{c_1} [\ldots v_1]) \in E(\mathrm{red}(G(S,\mu_b)))$*, men* $[v_2\ldots] \xrightarrow{c_2} [\ldots v_1]) \notin E(\mathrm{red}(G(S,\mu_b)))$ *då ger* $\mu_{b+1}(v_2) = v_1$ *en ändring med* 1 *i antalet udda cykler, det vill säga*

$$c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_b)),c_1,c_2) - c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_{b+1}))c_1,c_2) = \begin{cases} -1 & \textit{om cykeln är jämn} \\ +1 & \textit{om cykeln är udda} \end{cases}$$

Bevis. Med $u_1 = [\ldots v_1]$ och $u_2 = [v_2\ldots]$, blir den alternerande cykeln $(u_0 \xrightarrow{c_2} u_1 \xleftarrow{c_1} u_2 \xrightarrow{c_2} u_3 \xleftarrow{c_1} \ldots \xleftarrow{c_1} u_0)$ reducerad till $(u_0 \xrightarrow{c_2} u_3 \xleftarrow{c_1} \ldots \xleftarrow{c_1} u_0)$. Cykelns längd minskar med 1, vilket betyder att en jämn cykel blir udda och en udda cykel blir jämn. Övriga cykler påverkas inte. □

Lemma 3.0.7. *Låt* $u_1 = [\ldots v_1]$ *och* $u_2 = [v_2\ldots]$*. Antag att* $(u_0 \xrightarrow{c_1} u_1)$ *och* $(u_2 \xrightarrow{c_2} u_3)$*, där* $u_0 \neq u_2, u_1 \neq u_3$*, tillhör* $E(\mathrm{red}(G(S,\mu_b)))$*, vilket ger oss att ingen av* $(u_2 \xrightarrow{c_1} u_1)$ *och* $(u_2 \xrightarrow{c_2} u_1)$ *tillhör* $E(\mathrm{red}(G(S,\mu_b)))$*. Om vi då sätter* $\mu_{b+1}(v_2) = v_1$ *så följer det att* $c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_b)),c_1,c_2) - c_{\mathrm{u}}(\mathrm{red}(G(S,\mu_{b+1})),c_1,c_2) = \pm 1$

Bevis. Om den alternerande cykeln $(u_0 \xrightarrow{c_1} u_1 \xleftarrow{c_2} u_4 \xrightarrow{c_1} \cdots \xrightarrow{c_1} u_3 \xleftarrow{c_2} u_2 \xrightarrow{c_1} u_5 \xleftarrow{c_2} \cdots \xleftarrow{c_2} u_0)$ existerar, kommer den delas i två cykler, nämligen $(u_0 \xrightarrow{c_1} u_5 \xleftarrow{c_2} \cdots \xleftarrow{c_2} u_0)$ och $(u_4 \xrightarrow{c_2} u_3 \xleftarrow{c_1} \cdots \xleftarrow{c_1} u_4)$. Totallängden på cykeln minskar med 1. Detta kan ske på följande sätt:

1. En udda cykel delas till två udda eller två jämna. $\Rightarrow \pm 1$ udda cykel.

2. En jämn cykel delas till en jämn och en udda cykel. $\Rightarrow +1$ udda cykel.

Annars har vi de två alternerande cyklarna $(u_0 \xrightarrow{c_1} u_1 \xleftarrow{c_2} u_4 \xrightarrow{c_1} \cdots \xleftarrow{c_2} u_0)$ och $(u_5 \xleftarrow{c_1} u_2 \xrightarrow{c_2} u_3 \xleftarrow{c_1} \cdots \xrightarrow{c_2} u_3)$, som binds ihop till $(u_0 \xrightarrow{c_1} u_5 \xleftarrow{c_2} \cdots \xrightarrow{c_1} u_3 \xleftarrow{c_2} u_4 \xrightarrow{c_1} \cdots \xleftarrow{c_2} u_0)$. Totallängden på cykeln minskar med 1. Detta kan ske på tre olika sätt:

1. En udda och en jämn cykel blir en jämn cykel. $\Rightarrow -1$ udda cykel.

2. Två jämna cykler blir en udda cykel. $\Rightarrow +1$ udda cykel.

3. Två udda cykler blir en udda cykel. $\Rightarrow -1$ udda cykel.

Övriga alternerande cykler påverkas inte. $\square$

Lemma 3.0.8. *Antag att μ_b kan slutföras till alla medianer i $M(S)$. Om $|((v_2 \ldots) \longrightarrow (\ldots v_1))|_{\mathrm{red}(G(S,\mu_b))} = k$, så är $\mu(v_2) = v_1$ för vissa $\mu \in M(S)$.*

Bevis. Antag att μ är en median sådan att $\mu^{-1}(v_1) = v_3 \neq v_2$. Betrakta nu $\tau = \mu \circ (v_2 v_3)$, som uppfyller $\tau(v_2) = \mu(v_3) = v_1$. Eftersom transpositionen $(v_2 v_3)$ delar alternerande cykler i alla färger, så kommer antalet udda cykler inte att minska i någon färg. Därav följer $\delta(\tau, \mu) \leq \delta(\mu, S)$, så τ är en median. Då följer att antalet udda cykler med färgerna i och $k+1$ som uppfyller $(v_2 \ldots) \xrightarrow{i} (\ldots v_1)$ och som innehåller de svarta streck som transponeras ökar eller förblir oförändrat lokalt eftersom en jämn cykel delas till två udda cykler eller en udda cykel delas till en udda och en jämn. $\square$

Litteraturförteckning

[1] Niklas Eriksen: Median clouds and a fast transposition median solver, preprint.

[2] Niklas Eriksen, Axel Hultman: Estimating the expected reversal distance after a fixed number of reversals, *Advances of Applied Mathematics* **32** (2004), 439–453.

[3] Guillame Fertin, Anthony Labarre, Irena Rusu, Éric Tannier och Stéphane Vialette: *Combinatorics of genome rearrangements*. Massachusetts Institute of Technology, 2009.