

De afstand tus- sen ongewor- telde fylogenetische bomen

Margot van Aken

De afstand tussen ongewortelde fylogenetische bomen

door

Margot van Aken

ter verkrijging van de graad van Bachelor of Science
aan de Technische Universiteit Delft,
in het openbaar te verdedigen op woensdag 24 augustus om 11:00 uur.

Studentnummer:	4161785
Projectduur:	1 maart, 2016 – 25 augustus, 2016
Beoordelingscommissie:	Dr. Ir. L. van Iersel, TU Delft, supervisor Prof. Dr. Ir. A. W. Heemink, TU Delft Dr. B. van den Dries, TU Delft

Voorwoord

Het verslag dat voor u ligt is het resultaat van een vier maanden durend Bachelor Eind Project (BEP) ter afsluiting van de Bachelor Technische Wiskunde aan de TU Delft. Voor mijn BEP wilde ik graag een onderwerp in de sectie Optimalisering. Al snel werd ik gekoppeld aan mijn begeleider Leo van Iersel. Mijn opdracht was een ILP formulering te maken voor de vergelijking van fylogenetische bomen.

De eerste maand van mijn onderzoek heb ik me beziggehouden met het eigen maken van de theorie omtrent fylogenetische bomen. Hiervoor heb ik geschikte literatuur tot mijn beschikking gekregen van Leo van Iersel. Na de stof goed te beheersen begon het puzzelen naar een oplossing. Vragen die ontstonden tijdens mijn onderzoek werden beantwoord door Leo van Iersel. Een enkele vraag werd beantwoord door Catherine Mc-Cardin, die ik daarvoor dank.

Bij dezen wil ik graag een bijzonder woord van dank richten aan Leo van Iersel voor de fijne begeleiding en de tijd die hij vrijgemaakt heeft om iedere week met mij te zitten, mee te denken en al mijn vragen te beantwoorden.

Ik wens u veel leesplezier.

*Margot van Aken
Delft, juli 2016*

Inhoudsopgave

Inleiding en Probleemstelling	1
1 Afstand tussen binaire bomen	3
1.1 Fylogenetische bomen	3
1.2 Omvormen van fylogenetische bomen	3
1.3 Agreement Forest	4
2 Reduceren van bomen	7
3 Agreement forest van gewortelde bomen	9
3.1 Onverenigbare trippels	9
3.2 Restricties.	10
4 Agreement forests van ongewortelde fylogenetische bomen	11
4.1 Onverenigbare kwartetten	11
4.2 Fase één.	13
4.3 β -mapping	14
4.4 Fase twee	14
4.5 Algoritme voor Fase één en twee	15
5 ILP formulering voor exacte berekening van de TBR-afstand	17
5.1 Onverenigbare kwartetten	17
5.2 ILP formulering	18
Conclusie	18
Referenties	21

Inleiding en Probleemstelling

Een ongewortelde, binaire, fylogenetische boom is een boom waarvan de bladeren gelabeld zijn en iedere niet-gelabelde knoop graad 3 heeft. Fylogenetische bomen worden gebruikt om de ontstaansgeschiedenis van organismen te begrijpen. Twee fylogenetische bomen T_1 en T_2 met dezelfde gelabelde blad-verzameling $\mathcal{L}$ kunnen namelijk vergeleken worden. Kan het ene organisme ontstaan zijn uit het andere organisme? Zo ja, hoe groot is de vergelijking? Door middel van zogenoemde Subtree Transfer-operaties kan T_1 omgevormd worden tot T_2 . De twee belangrijkste Subtree Transfer-operaties in dit verslag zijn SPR-operaties en TBR-operaties. Bij beide operaties wordt er eerst een tak verwijderd van de boom. Bij een SPR-operatie wordt een tak geknipt en aan een andere tak vastgemaakt, terwijl bij een TBR-operatie een tak volledig verwijderd wordt, waarna een nieuwe tak de verbinding zal zijn tussen de ontstane deelbomen. Hoe vaak een operatie moet worden toegepast om T_1 om te vormen tot T_2 heet de afstand tussen de bomen. De probleemstelling van dit verslag luidt: hoe kan de TBR-afstand tussen twee fylogenetische bomen bepaald worden?

Het blijkt dat de exacte afstand tussen T_1 en T_2 verbonden is met het aantal componenten in een Maximum Agreement Forest van T_1 en T_2 [1]. Een MAF kan omschreven worden als het minimale aantal bomen, zó dat iedere boom een deelboom van T_1 en T_2 is; de bomen disjunct zijn; en deze bomen heel $\mathcal{L}$ opspannen.

Aangezien voor de probleemstelling een MAF gevonden moet worden tussen twee fylogenetische bomen, wordt in hoofdstuk 2 een handig uitstapje gemaakt. Wanneer twee bomen topologisch verwant zijn, kunnen ze gereduceerd worden, zonder dat de TBR-afstand verandert [1].

In hoofdstuk 3 wordt laten zien dat bij gewortelde, binaire bomen gekeken wordt naar onverenigbare trippels om een agreement forest te vinden. Twee gewortelde, binaire bomen zijn namelijk equivalent dan en slechts dan als iedere trippel in de twee bomen topologisch equivalent is [2].

Bij twee fylogenetische bomen T_1 en T_2 wordt er gekeken naar onverenigbare kwartetten. In hoofdstuk 4 wordt een algoritme beschreven die de TBR-afstand berekent [3]. Deze berekening bestaat uit twee fasen. In de eerste fase worden van bepaalde onverenigbare kwartetten in T_2 een tak verwijderd, zodat er een bos $\mathcal{F}'$ overblijft. In de tweede fase wordt gekeken naar paden die disjunct zijn in $\mathcal{F}'$, maar overlappen in T_2 . Weer worden er takken verwijderd, zodat na de tweede fase een agreement forest overblijft.

Tot slot wordt in het laatste hoofdstuk de theorie van hoofdstuk 3 en 4 gecombineerd. Deze combinatie resulteert in een nieuwe theorie. Dit geeft een ILP-formulering waarin de grootte van een MAF tussen twee fylogenetische bomen gevonden kan worden.

Er is in dit verslag gekozen om alleen de belangrijkste lemma's en stellingen te bewijzen. De kern van dit verslag ligt in de laatste drie hoofdstukken, dus daar zullen meer bewijzen te vinden zijn dan in de eerste twee.

Afstand tussen binaire bomen

Dit hoofdstuk zal een fundering leggen voor de kennis die nodig is voor dit verdere verslag, gebaseerd op een artikel van Allen en Steel [1]. Eerst zullen een aantal definities gegeven worden omtrent fylogenetische bomen. In de tweede paragraaf zal er uitgelegd worden op welke manieren bomen omgevormd kunnen worden. In hoeveel stappen van een boom T_1 een boom T_2 gemaakt kan worden, noemen we de afstand tussen deze bomen. Hiervoor gelden verschillende afstandmaten die gebaseerd zijn op verschillende regels voor het omvormen. In de laatste paragraaf van dit hoofdstuk wordt het verband uitgelegd tussen deze afstand en een zogeheten Maximum Agreement Forest. Dit verband blijkt nuttig te zijn wanneer we de afstanden willen berekenen.

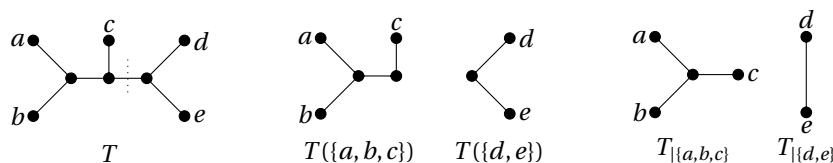
1.1. Fylogenetische bomen

Een *ongewortelde, binaire, fylogenetische boom* is een boom waarvan de bladeren gelabeld zijn en iedere niet-gelabelde knoop graad 3 heeft. Zulke bomen noemen we ook wel kortweg *fylogenetische bomen*. De verzameling fylogenetische bomen met n bladeren wordt aangeduid met $UB(n)$. De verzameling gelabelde bladeren van een boom T wordt genoteerd als $\mathcal{L}(T)$.

Een tak die niet verbonden is met een blad, noemen we een *interne tak*. Een interne tak is altijd verbonden aan 4 deelbomen.

Een boom met punten van graad 2 kan binair gemaakt worden door *forced contraction*. Dit is een operatie waarbij knopen van graad twee verwijderd worden en de aan deze knoop liggende takken samengevoegd worden.

Gegeven een verzameling $U \subseteq \mathcal{L}(T)$ voor een binaire boom T , laat dan $T(U)$ de minimale deelboom van T zijn die alle bladeren van U verbindt. We noteren dan $T|_U$ voor de binaire boom verkregen uit $T(U)$ door forced contraction. Zie Figuur 1.1 voor een voorbeeld.

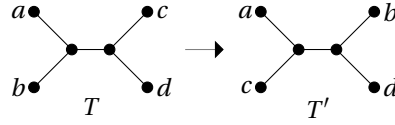


Figuur 1.1: Schematische weergave van forced contraction.

1.2. Omvormen van fylogenetische bomen

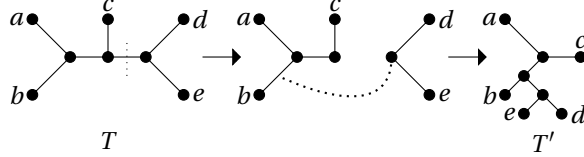
Twee binaire bomen met een gelijk aantal bladeren kunnen zó omgevormd worden dat de bomen hetzelfde (isomorf) worden. Een boom omvormen kan aan de hand van een *Subtree Transfer-operatie*. We maken onderscheid tussen drie soorten Subtree Transfer-operaties (Θ -operaties), namelijk Nearest Neighbour Interchange (NNI), Subtree Prune and Regraft (SPR) en Tree Bisection and Reconnection (TBR).

Bij een *NNI-operatie* wordt een deelboom aansluitend aan een interne tak verwisseld met een deelboom aan de andere kant van de interne tak, zoals te zien in Figuur 1.2.



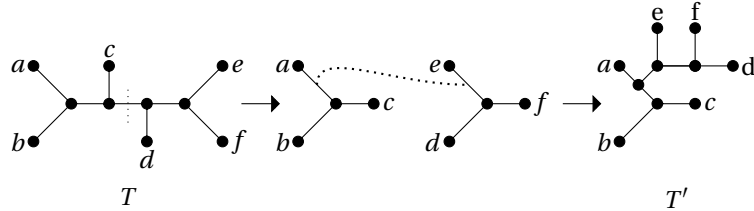
Figuur 1.2: Schematische weergave van een NNI-operatie.

Bij een *SPR-operatie* wordt er een tak verwijderd. De ontstane deelboom wordt vervolgens aan diezelfde tak aan een ander deel van de boom vastgemaakt. Met forced contraction wordt de boom weer binair. Zie Figuur 1.3 voor een visuele uitleg.



Figuur 1.3: Schematische weergave van een SPR-operatie.

Als laatste wordt er bij een *TBR-operatie* een tak verwijderd en worden de twee ontstane deelbomen samengevoegd door een tak toe te voegen zoals te zien is in Figuur 1.4. Weer zorgt forced contraction voor de binairiteit.



Figuur 1.4: Schematische weergave van een TBR-operatie.

Merk op dat elke NNI-operatie ook een SPR-operatie is en elke SPR-operatie ook een TBR-operatie is, maar andersom dit niet altijd geldt.

De *afstand* tussen twee binaire bomen T_1 en T_2 ten opzichte van een operatie $\Theta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$ is het minimale aantal Θ -operaties dat nodig is om T_2 uit T_1 te verkrijgen. We noteren dit als $d_\Theta(T_1, T_2)$.

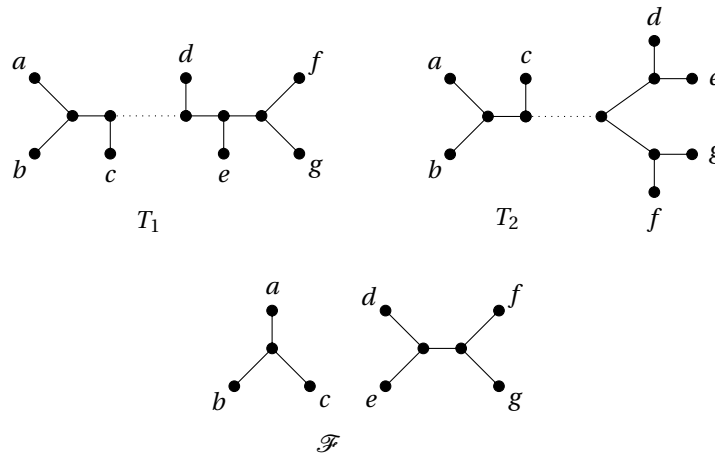
1.3. Agreement Forest

De TBR-afstand tussen twee binaire bomen T_1 en T_2 kan aan de hand van een *maximum agreement forest* (MAF) bepaald worden. Het verband tussen de MAF en de afstand wordt aan het einde van deze paragraaf verduidelijkt. Een MAF voor bomen $T_1, T_2 \in \text{UB}(n)$ kan omschreven worden als het minimale aantal bomen, zodat iedere boom een deelboom van T_1 en T_2 is; de bomen disjunct zijn; en deze bomen heel $\mathcal{L}$ opspannen. De definitie volgt.

Laat $\mathcal{L} := \mathcal{L}(T_1) = \mathcal{L}(T_2)$. Een *agreement forest* (AF) van T_1 en T_2 is een verzameling $\mathcal{F} = \{t_1, \dots, t_k\}$ binaire bomen zodat als we $\mathcal{L}_j := \mathcal{L}(t_j)$ voor $j \in \{1, \dots, k\}$ laten, dan:

1. $\mathcal{L}_1, \dots, \mathcal{L}_k \subseteq \mathcal{L}$,
2. $t_j = T_1|_{\mathcal{L}_j} = T_2|_{\mathcal{L}_j}$ voor alle $j \in \{1, \dots, k\}$, en
3. Voor $i \in \{1, 2\}$ zijn de bomen $\{T_i(\mathcal{L}_j) : j = 1, \dots, k\}$ knoop-disjunct.

Een *maximum agreement forest* (MAF) is een AF $\mathcal{F}$ waarbij $|\mathcal{F}|$ minimaal is, zie Figuur 1.4. Merk op dat een MAF niet uniek hoeft te zijn.



Figuur 1.4: Een MAF $\mathcal{F}$ voor $T_1, T_2 \in UB(n)$.

Een MAF bestaat uit zo min mogelijk deelbomen. We kunnen de MAF van twee bomen gebruiken om hun afstand te bepalen, door middel van Stelling 1.

Stelling 1. Voor twee binaire bomen T_1 en T_2 met $\mathcal{L}(T_1) = \mathcal{L}(T_2)$ geldt

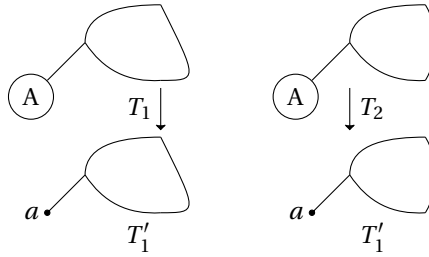
$$d_{TBR}(T_1, T_2) = |\text{MAF}(T_1, T_2)| - 1 \quad \text{en} \quad d_{SPR}(T_1, T_2) = |\text{MAF}| - 1.$$

2

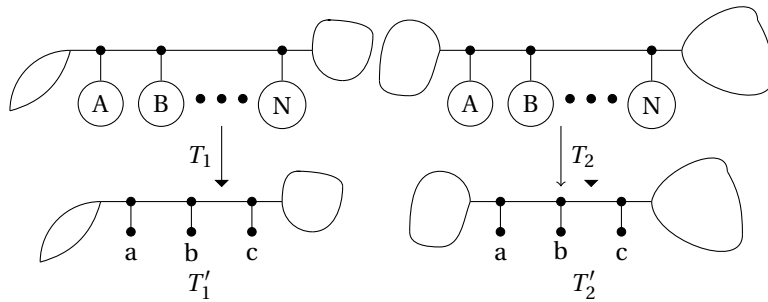
Reduceren van bomen

Nu bekend is dat de afstand tussen twee bomen bepaald kan worden aan de hand van de MAF, is het zaak de MAF te vinden. De eerste stap die gemaakt moet worden is het reduceren van de bomen tot zo klein mogelijke bomen. Dit kan gedaan worden aan de hand van het herhaaldelijk toepassen van de volgende regels:

1. Vervang identieke deelbomen door een blad.
2. Vervang identieke ketens van deelbomen door 3 bladeren.



Figuur 2.1: Schematische weergave van Regel 1.



Figuur 2.2: Schematische weergave van Regel 2.

Een *abc-boom* is een binaire boom T die bladeren a, b en c heeft die respectievelijk aan knopen v_a, v_b en v_c vastzitten, zó dat $\{v_a, v_b\}$ en $\{v_b, v_c\}$ takken van T zijn. De bomen in Figuur 2.2 zijn hier voorbeelden van.

Lemma 1. Als $T_1, T_2 \in UB(n)$ twee *abc-bomen* zijn met $\mathcal{L}(T_1) = \mathcal{L}(T_2)$, dan bestaat er een MAF $\mathcal{F}$ voor T_1, T_2 waarin a, b, c samenhangend zijn in een van de bomen van $\mathcal{F}$.

Het blijkt dat bomen $T_1, T_2 \in UB(n)$ in polynomiale tijd maximaal gereduceerd kunnen worden met behulp van Regels 1 en 2. Dit reduceren kan worden gebruikt om $d_{TBR}(T_1, T_2)$ sneller te bepalen, zoals de volgende stelling zegt. Deze kan worden bewezen met behulp van Lemma 1.

Stelling 2. *Laat $T_1, T_2 \in UB(n)$ en T'_1 en T'_2 de gereduceerde bomen van respectievelijk T_1 en T_2 . Dan $d_{TBR}(T_1, T_2) = d_{TBR}(T'_1, T'_2)$.*

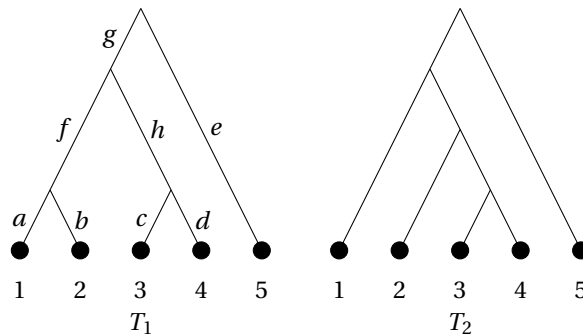
Deze stelling geldt niet voor het NNI-afstand probleem, aangezien voor regel (2) niet afstand-behoudend is. Voor de SPR-afstand is zover nog niet bekend of de stelling van toepassing is.

Agreement forest van gewortelde bomen

De begrippen SPR-afstand en TBR-afstand zijn nauw verbonden. Er is een sterke relatie tussen de SPR-afstand en een MAF van gewortelde, binaire bomen, terwijl voor ongewortelde, binaire bomen die sterke relatie bestaat tussen een MAF en de TBR-afstand. Omdat er een parallel te vinden is tussen de berekening voor een SPR-afstand en de berekening voor een TBR-afstand, worden hier een aantal belangrijke begrippen en stellingen omtrent de SPR-afstand toegelicht.

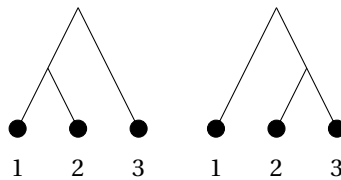
3.1. Onverenigbare trippels

De verzameling van binaire, gewortelde bomen met n bladeren noteren we als $RB(n)$. Er zullen een aantal definities worden uitgelegd aan de hand van onderstaande figuren.



Figuur 3.1: Een illustratie van twee gewortelde, binaire bomen T_1 en T_2 .

Een geïnduceerde deelboom met een drietal bladeren noemen we een *trippel*. We bekijken trippels behorende bij de bladeren 1, 2 en 3.



Figuur 3.2: trippels van bladeren 1, 2, 3 behorende bij resp. T_1 en T_2 .

In de bovenstaande figuur zijn twee trippels te zien die niet gelijk zijn. We noemen een dergelijke trippel een *onverenigbare trippel*. Een *wandeling* w in een boom T is een aaneenschakeling van knooppunten $w = [v_1, v_2, \dots, v_j]$ in T , zodanig dat $\{v_i, v_{i+1}\} \in E(T)$ voor alle $j = 1, \dots, i - 1$. Een wandeling zonder repetities van knopen is een *pad*. We noteren het pad tussen twee bladeren 1 en 2 als $Pad(1, 2)$. Laat nu 1, 2 en 3

bladeren van een trippel zijn. We definiëren $TriWand(1,2,3) := Pad(1,2) \cap Pad(1,3) \cap Pad(2,3)$. Ter voorbeeld schrijven we voor bladeren 1 en 3 van T_1 dus $Pad(1,3) = \{a, c, f, h\}$ en voor bladeren 1,2 en 3 schrijven we $TriWand(\{1,2,3\}) = \{a, b, c, f, h\}$.

Lemma 2. *Twee gewortelde, binaire bomen zijn equivalent dan en slechts dan als iedere trippel van bladeren in de twee bomen equivalent is.*

3.2. Restricties

Omdat er een sterke relatie is tussen de SPR-afstand en een MAF van gewortelde, binarie bomen, willen we een MAF vinden. Volgens bovenstaand lemma zal in een MAF dus geen onverenigbaar trippel zitten. We definiëren twee types restricties voor het verwijderen van takken:

- *Type één:* Stel de trippel i, j, k van T_1 en T_2 is onverenigbaar. Dan moet er tenminste één tak van $T_1|_{\{i,j,k\}}$ worden verwijderd.
- *Type twee:* Stel voor 4 bladeren i, j, k en l geldt dat $Pad(i, j) \cap Pad(k, l) \neq \emptyset$ in T_2 , maar $Pad(i, j) \cap Pad(k, l) = \emptyset$ in T_1 , dan moet er tenminste één van de takken van $Pad(i, j) \cup Pad(k, l)$ in T_1 worden verwijderd.

Lemma 3. *Laat $T_1, T_2 \in RB(n)$. Laat $\mathcal{F} = \{t_1, \dots, t_k\}$ een bos verkregen uit T_1 door $T_1 \setminus F$, met $F \subset E(T_1)$. Dan zijn de volgende uitspraken equivalent:*

1. $\mathcal{F}$ is een agreement forest.
2. $e_i \in F$ voldoet aan restricties type één en twee $\forall i \in \{1, \dots, n\}$.

Bewijs. Uit de definitie van een agreement forest volgt dat t_j in T_2 zit $\forall j \in \{1, \dots, k\}$ en dat de bomen bovendien onderling knoop-disjunct zijn in T_2 . Aan type één wordt dus voldaan. Er kan immers geen onverenigbare trippel in $\mathcal{F}$ zitten, aangezien de bomen van $\mathcal{F}$ zowel in T_1 als in T_2 zitten. Iedere trippel van drie bladeren uit dezelfde component van $\mathcal{F}$ is namelijk verenigbaar met beide bomen.

Voor restrictie type twee gaan we een tegenspraak afleiden. Neem aan dat voor een viertal bladeren i, j, k en l geldt dat $Pad(i, j) \cap Pad(k, l) \neq \emptyset$ in T_2 , maar de paden wel disjunct zijn in T_1 en er geen tak is verwijderd van de genoemde paden. Dan moet er dus gelden dat i, j in één boom zitten en k, l in één boom zitten. Omdat $Pad(i, j) \cap Pad(k, l) \neq \emptyset$ in T_2 geldt er dan zelfs dat i, j, k en l in dezelfde boom van $\mathcal{F}$ zitten. Dit betekent echter dat er een onverenigbare trippel in $\mathcal{F}$ zit, waarvan we net hebben laten zien dat dit niet mogelijk is. We concluderen dat het verwijderen voldoet aan restricties type één en type twee.

Anderzijds, neem nu aan dat bepaalde takken van T_1 verwijderd zijn onder voorwaarden van restricties type één en type twee, zó dat er een bos $\mathcal{F} = \{t_1, \dots, t_k\}$ is ontstaan. We gaan laten zien dat $\mathcal{F}$ een agreement forest is. Dit doen we aan de hand van de definitie.

1. Duidelijk dat $\mathcal{L}_1, \dots, \mathcal{L}_k \subseteq \mathcal{L}$.
2. Door restrictie type één zitten er geen onverenigbare trippels van T_1 en T_2 in $\mathcal{F}$. Nu volgt uit Lemma 2 dat $\forall j \in \{1, \dots, k\}$ de boom $t_j \in \mathcal{F}$ een geïnduceerde deelboom is van T_2 . Dus $t_j = T_1|_{\mathcal{L}_j} = T_2|_{\mathcal{L}_j}$ voor alle $j \in \{1, \dots, k\}$.
3. We laten zien dat voor $i \in \{1, 2\}$ de bomen $\{T_1(\mathcal{L}_j) : j = 1, \dots, k\}$ knoop-disjunct zijn, door tegenspraak af te leiden. Stel t_n, t_m in T_2 zijn niet pad-disjunct voor zekere $n, m \in \{1, \dots, l\}$ met $n \neq m$. Merk op dat t_n, t_m niet uit maar één blad kunnen bestaan, aangezien ze dan wel degelijk pad-disjunct zouden zijn. Kies nu bladeren $i, j \in t_n$ en $k, l \in t_m$ zó dat $Pad(i, j) \cap Pad(k, l) \neq \emptyset$ in T_2 . Omdat $i, j \in t_n$ en $k, l \in t_m$, zijn de paden wél disjunct in T_1 . Door restrictie type twee moet er echter wel degelijk een tak zijn verwijderd van $Pad(i, j) \cup Pad(k, l)$. Maar dan $i, j \notin t_n$ óf $k, l \notin t_m$. Dit is een tegenspraak, dus t_n, t_m zijn pad-disjunct in T_2 . Omdat T_2 binair is, volgt dat t_n, t_m knoop-disjunct zijn.

We concluderen dat $\mathcal{F}$ een agreement forest is. □

Uit Lemma 3 volgt dat een MAF gevonden kan worden door zo min mogelijk takken te verwijderen, zodanig dat aan restrictie type één en twee voldaan wordt.

Agreement forests van ongewortelde fylogenetische bomen

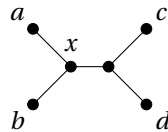
Dit hoofdstuk beschrijft een algoritme voor het vinden van een agreement forest voor twee ongewortelde fylogenetische bomen T_1 en T_2 . Dit algoritme resulteert uit een onderzoek van Hallet en McCartin (2007) [3] en is fixed-parameter tractable. Dat wil zeggen dat het algoritme een agreement forest kan vinden binnen een bepaalde polynomiale tijd. Het algoritme bestaat uit twee fasen. Om aan voorwaarden (i) en (ii) van de definitie van een agreement forest te voldoen, zullen er eerst bepaalde takken verwijderd moeten worden van onverenigbare kwartetten. Hierna zullen in fase twee zogenoemde *obstructies* verwijderd worden, zodat ook aan de voorwaarde (iii) wordt voldaan.

4.1. Onverenigbare kwartetten

Bij gewortelde bomen kan een agreement forest gevonden worden aan de hand van trippels. Voor een agreement forest van twee ongewortelde fylogenetische bomen zal er gekeken worden naar kwartetten. Laat T_1 een fylogenetische boom zijn. Voor drie knopen $u_1, u_2, u_3 \in V(T_1)$, definiëren we

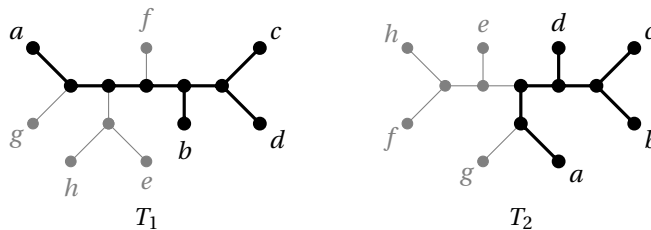
$$P_F(u_1, u_2, u_3) = \{x \in V(T_1) : x \in \bigcap_{1 \leq i < j \leq 3} \text{Pad}(u_i, u_j)\}.$$

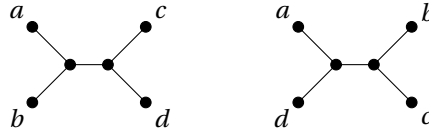
Voor een viertal knopen $a, b, c, d \in \mathcal{L}(T_1)$ waarvan $T_1[\{a, b, c, d\}]$ samenhangend is, noemen we $Q = T_1|_K$ met $K = \{a, b, c, d\}$ een *kwartet*. Zijn a en b broers (en c en d dus ook), dan schrijven we kortweg $ab|cd$.



Figuur 4.1: Kwartet $ab|cd$ met $x = P_F(a, b, c)$.

Laat nu T_2 ook een fylogenetische boom zijn met $\mathcal{L}(T_1) = \mathcal{L}(T_2)$. We noemen Q *verenigbaar* met T_2 dan en slechts dan als $T_2|_K$ en Q isomorf zijn. Anders noemen we het kwartet *onverenigbaar*.

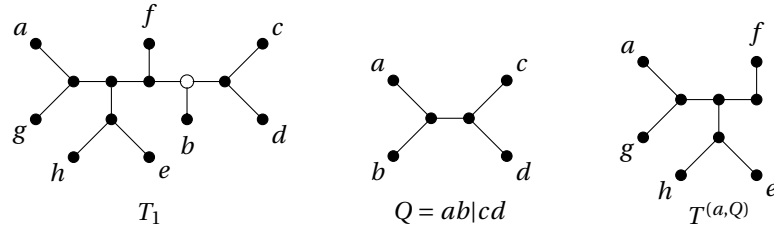




Figuur 4.2: Fylogenetische bomen T_1 en T_2 met een onverenigbaar kwartet.

Merk op dat het kwartet $ab|cd$ in T_1 verenigbaar is met T_2 dan en slechts dan als $Pad(a, b) \cap Pad(c, d) = \emptyset$ in T_2 . We noemen twee fylogenetische bossen F_1 en F_2 *verenigbaar* wanneer ieder kwartet in F_1 verenigbaar is met F_2 , óf onafhankelijk is in F_2 .

Voor een kwartet $Q = ab|cd$ in een boom T , noemen we het punt $e \in V(T)$ waarvoor a, b en c in drie verschillende deelbomen van $T \setminus e$ zitten het *verbindingspunt* van a of b in T . We noteren het samenhangende component in $T \setminus e$ die a bevat als $T^{(a,Q)}$, zie figuur 4.3. We noemen $l \in \mathcal{L}(T^{(a,Q)})$ met $l \neq a$ de *zusters* van a in $(T^{(a,Q)})$.

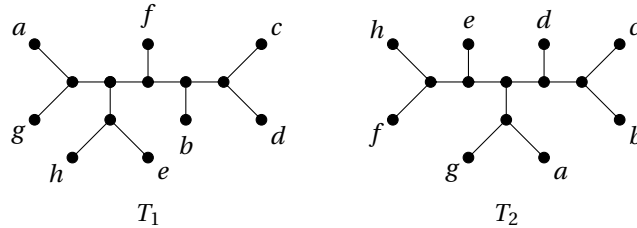


Figuur 4.3: Fylogenetische boom T , met kwartet $Q = ab|cd$ en deelboom $T^{(a,Q)}$.

We introduceren een nieuwe definitie. Laat een kwartet $Q = ab|cd$ in T_1 onverenigbaar zijn met T_2 , dan is het een *minimaal overenigbaar kwartet* (MOK) als hij voldoet aan één van de volgende eisen:

1. voor alle $x \in \{a, b, c, d\}$ is het nieuwe kwartet Q' , gevormd door de broer van x in Q te vervangen door een zuster van x in $(T^{(x,Q)})$ wél verenigbaar met T_2 .
2. geen $x \in \{a, b, c, d\}$ heeft zusters in Q .

In andere woorden heeft een MOK dus het langst mogelijke interne pad. Wanneer het pad langer wordt, zal het kwartet verenigbaar worden, zie figuur 4.4.

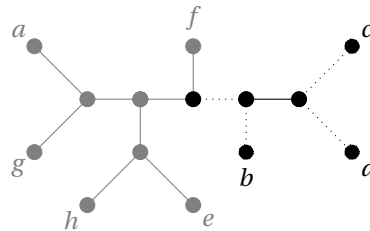


Figuur 4.4: $ab|cd$ is een MOK, want $ae|cd, af|cd, ag|cd$ en $ah|cd$ zijn verenigbare kwartetten.

Voor $Q = ab|cd$, waarbij n het verbindingspunt van a en m het verbindingspunt is van c , definiëren we de *vork* R geïnduceerd door Q in T_1 als

$$T_1[\{x \in V(T_1) : x \in Pad(n, m) \text{ of } x \text{ is aangrenzend aan } n \text{ of } m\}]$$

De takken aan de uiteindes van de vork noemen we de *tanden* van de vork, zie figuur 4.5.



Figuur 4.5: Fylogenetische boom T , de vork van kwartet $Q = ab|cd$ met de tanden gestippeld.

Merk op dat dan $Q' = xb|cd$ met x een zuster van a in $\mathcal{L}(T^{(a,Q)})$ ook een MOK is en dat de vork van Q gelijk is aan de vork van Q' .

We zijn aangekomen bij een belangrijk lemma, welke als volgt geïnterpreteerd kan worden. Voor twee bomen $T_1, T_2 \in UB(n)$ kan een agreement forrest verkregen worden door te kijken naar Minimaal Onverenigbare Kwartetten. Stel $Q = ab|cd$ is een MOK in T_1 van T_2 , dan moet tenminste één van de tanden van Q verwijderd worden om een agreement forrest te verkrijgen. Het lemma volgt.

Lemma 4. *Laat $T_1, T_2 \in UB(n)$ en laat $Q = ab|cd$ een MOK in T_1 . In een agreement forrest $\mathcal{F} = \{t_1, \dots, t_k\}$ van T_1 en T_2 geldt $\{l_a, l_b, l_c, l_d\} \not\subseteq \mathcal{L}(t_i)$ voor alle $l_x \in T_1^{(x,Q)}$, $x \in \{a, b, c, d\}$ en $1 \leq i \leq k$.*

Bewijs. Laat $\mathcal{F} = \{t_1, \dots, t_k\}$ een AF van T_1 en T_2 zijn. Laat $Q = ab|cd$ een MOK in T_1 . Kies nu knopen $l_x \in T_1^{(x,Q)}$ voor alle $x \in \{a, b, c, d\}$. Laat $C = \{l_a, l_b, l_c, l_d\}$. Stel nu dat er een $t_i \in \mathcal{F}$ bestaat met $C \subseteq \mathcal{L}(t_i)$. We gaan tegenspraak afleiden.

Noem $Q_C = t_i|C$. Omdat $\mathcal{F}$ een agreement forest is, volgt dat Q_C , $T_1|C$ en $T_2|C$ isomorf zijn.

Omdat Q een MOK is en dus onverenigbaar, volgt dat Q_C en Q niet isomorf zijn. Er moet dus voor tenminste één $x \in \{a, b, c, d\}$ gelden dat $l_x \neq x$.

Aangezien T_2 een fylogenetische boom is, moet er gelden dat ofwel (i) $ac|bd$ ofwel (ii) $ad|bc$ verenigbaar is met T_2 . We gaan nu gebruik maken van de minimaliteit van Q . Er moet nu namelijk gelden dat $a = l_a$ óf $al_a|ef$ is verenigbaar met T_2 , $b = l_b$ óf $bl_b|ef$ is verenigbaar met T_2 , $c = l_c$ óf $ab|cl_c$ is verenigbaar met T_2 en $d = l_d$ óf $ab|dl_d$ is verenigbaar met T_2 .

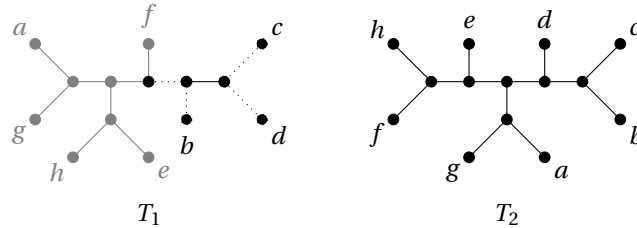
Hierdoor moet er in geval (i) gelden dat $l_al_c|l_b|l_d$ verenigbaar is met T_2 . Dit is in tegenspraak met de verenigbaarheid van Q_C , aangezien $l_al_c|l_b|l_d \neq l_al_b|l_cl_d = Q_C$.

Op dezelfde manier geldt in geval (ii) dan dat $l_al_d|l_b|l_c$ verenigbaar is met T_2 . Dit geeft weer tegenspraak met de verenigbaarheid van Q_C , aangezien $l_al_d|l_b|l_c \neq l_al_b|l_cl_d = Q_C$.

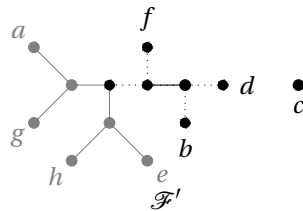
Een dergelijke verzameling C bestaat dus niet. We concluderen dat $\{l_a, l_b, l_c, l_d\} \not\subseteq \mathcal{L}(t_i)$ voor alle $l_x \in T_1^{(x,Q)}$, $x \in \{a, b, c, d\}$ en $1 \leq i \leq k$ $\square$

4.2. Fase één

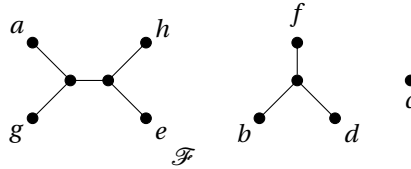
Aan de hand van Lemma 4 zal Fase één worden uitgevoerd. In Fase één wordt van een MOK in T_2 van T_1 één tand verwijderd. Deze stap wordt herhaald tot er een bos $\mathcal{F}$ ontstaan is, zó dat alle bomen in $\mathcal{F}$ geïnduceerde deelbomen van T_2 zijn en er geen MOK in $\mathcal{F}$ zit. Er volgt een voorbeeld. Weer nemen we $T_1, T_2 \in UB(n)$ zoals eerder dit hoofdstuk.



Eerder zagen we al dat $Q = ab|cd$ een MOK is. We verwijderen de tand naar c .



Nog een MOK die we in $\mathcal{F}'$ vinden is $eh|af$. We kiezen nu de tand naar f om te verwijderen.

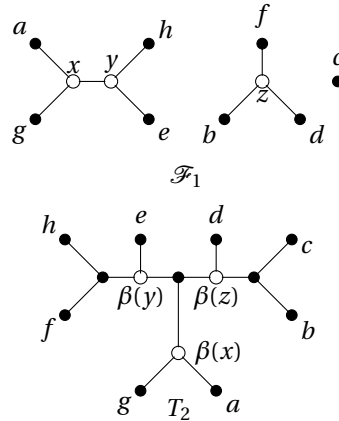


We hebben nu een bos $\mathcal{F}$ waarin geen MOK meer zit. Het kwartet $ag|eh$ is verenigbaar. Fase één is nu afgerond.

4.3. β -mapping

In Fase twee zullen weer takken verwijderd worden. Laat hiervoor het bos $\mathcal{F}' = \{t_1, \dots, t_k\}$ ontstaan zijn uit T_1 door het toepassen van Fase één. Duidelijk is dat $\mathcal{F}$ dan aan de eerste en tweede conditie voldoet van de definitie agreement forest. De bomen van $\mathcal{F}$ hoeven echter nog niet disjunct te zijn in T_2 . Om ook aan de derde conditie te voldoen introduceren we in deze paragraaf β -mapping en *obstructies*.

We definiëren β -mapping, waarbij voor ieder niet-gelabelde knoop u van $\mathcal{F}'$ een bijbehorende knoop in T_2 een β -label krijgt. Laat u het verbindingspunt zijn van componenten t_i, t_j en t_k in $\mathcal{F}_1$. Er geldt $\beta(u) = P_{T_2}(L(t_i), L(t_j), L(t_k))$, zie Figuur 4.6.



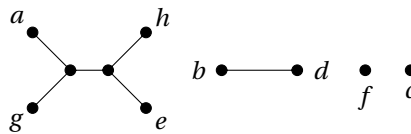
Figuur 4.6: Bos $\mathcal{F}_1$ en fylogenetische boom T_2 met β -mapping.

Laat (h, y) en (f, z) twee disjuncte takken zijn in $\mathcal{F}_1$. Als er een $x \in V(T_2)$ bestaat waarvoor geldt dat $x \in Pad(h, \beta(y)) \cap Pad(f, \beta(z))$, dan noemen we (h, y) en (f, z) *obstructies*. In twee obstructies moet tenminste één tak verwijderd worden in die obstructies om een agreement forest te verkrijgen. Dit leidt tot Fase twee.

4.4. Fase twee

Net als bij Fase één is er bij Fase twee een stap die herhaald wordt tot hij niet meer herhaald kan worden. Bij Fase twee draait het om de obstructies. Voor twee obstructies (a, b) en (c, d) in $\mathcal{F}'$ moet tenminste één van deze takken verwijderd worden. Deze stap wordt herhaald tot er geen obstructies meer zijn. Het ontstane bos is een agreement forest.

In het reeds genoemde voorbeeld hebben we obstructies (h, y) en (f, z) . We kiezen de tak (f, z) om te verwijderen:



We hebben nu een agreement forest. Merk op dat als (h, y) gekozen was om te verwijderen, opnieuw β -mapping toegepast moest worden om nog een obstructie te verwijderen. Door onfortuinlijk te kiezen kunnen

de stappen van zowel Fase één als van twee vaker dan nodig gemaakt moeten worden. Door Fase één en twee wordt dan ook een agreement forest verkregen, dat niet een MAF hoeft te zijn.

4.5. Algoritme voor Fase één en twee

Het probleem dat we op willen lossen is het vinden van een MAF. Door Fase één en twee uit te voeren ontstaat een agreement forest. Bij iedere takverwijdering de juiste tak kiezen zal echter tot een MAF leiden.

Zoals in Hoofdstuk 2 uitgelegd is, kunnen bomen in polynomiale tijd gereduceerd worden, zó dat de TBR-afstand tussen de bomen gelijk blijft. De bomen worden verkleind, waardoor er minder mogelijkheden zijn voor takverwijderingen.

In deze paragraaf staat een optimalisatie algoritme, dat het aantal takverwijderingen vermindert. De input bestaat uit twee gereduceerde fylogenetische bomen en een getal k waarvoor getest wordt of er een agreement forest bestaat met k componenten. Bestaat deze niet, dan zal k met 1 vergroot worden en het algoritme opnieuw worden uitgevoerd. Het algoritme ziet er als volgt uit:

Fase 1:

Input: T_2, F .

1. Als er geen MOK Q in F is en
 - (a) fase twee levert geen agreement forest F' op met $|F'| \leq k$, dan $F_{uit} = \text{null}$, stop.
 - (b) fase twee levert een agreement forest F' op met $|F'| \leq k$, dan $F_{uit} = F'$, stop.
2. Neem MOK Q met tanden $e_i, i = 1, \dots, 4$. Construeer de bossen $F_i = F \setminus e_i$ voor $i = 1, \dots, 4$.
3. Als één F_i een agreement forest is, dan $F_{uit} = F_i$. Anders reduceer F_i en vind F_{uit_i} voor elke F_i en T_2 recursief. Als één F_{uit_i} een agreement forest is, dan $F_{uit} = F_{uit_i}$.

Output: F_{uit} .

Fase 2:

Input: F, T_2 .

1. Als er geen obstructies e_n en e_m zijn in F , dan $F_{uit} = F$, stop.
2. Neem obstructies e_n, e_m . Construeer bossen $F_i = F \setminus e_i$ voor $i = n, m$.
3. Als $|F_i| \leq k$ vind F_{uit_i} voor elke F_i en T_2 recursief, anders stop.

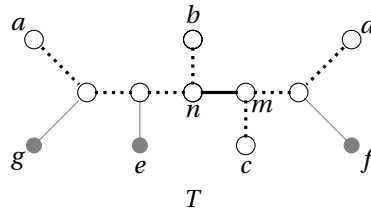
Output: F_{uit} .

ILP formulering voor exacte berekening van de TBR-afstand

In hoofdstuk 4 zagen we een manier om een MAF te vinden van twee fylogenetische bomen. De theorie is echter vrij lastig en het algoritme erg langzaam. In dit hoofdstuk wordt de methode van hoofdstuk 3 aangepast voor ongewortelde fylogenetische bomen, zodat een simpele ILP formulering gebruikt kan worden om het TBR-afstand probleem op te lossen.

5.1. Onverenigbare kwartetten

Voor het kwartet $Q = ab|cd$ in T_1 , waarbij n het verbindingspunt van a en m het verbindingspunt is van c , definiëren we de *armen* van Q als $A_Q := \{e \in E(T_1) : e \in \text{Pad}(a, b) \cup \text{Pad}(c, d)\}$ en de *ruggengraat* van Q als $G_Q := \{e \in E(T_1) : e \in \text{Pad}(n, m)\}$. Zie Figuur 5.1.



Figuur 5.1: Fylogenetische boom T , kwartet $Q = ab|cd$, met A_Q gestippeld en G_Q dikgedrukt.

Lemma 5. *Laat $T_1, T_2 \in UB(n)$. Laat $\mathcal{F} = \{t_1, \dots, t_k\}$ een bos verkregen door $T_1 \setminus \mathcal{E}(T_1)$, met $\mathcal{E}(T_1) \subset E(T_1)$. Dan zijn de volgende uitspraken equivalent:*

1. $\mathcal{F}$ is een *agreement forest*.
2. Voor ieder kwartet Q in T_1 dat onverenigbaar is met T_2 er tenminste één tak van A_Q verwijderd is in T_1 .

Bewijs. Laat $\mathcal{F} = \{t_1, \dots, t_k\}$ een agreement forest van T_1 en T_2 zijn, verkregen door een stel takken in T_1 te verwijderen. Laat $Q = ab|cd$ een onverenigbaar kwartet in T_1 zijn met T_2 . We onderscheiden twee gevallen:

1. We verwijderen geen tak van Q . Dan is $Q \in \mathcal{F}$. We hebben een tegenspraak, aangezien per definitie geen onverenigbaar kwartet in een agreement forest kan zitten.
2. We verwijderen een tak van de ruggengraat van Q . Q is onverenigbaar met T_2 , dus zit óf $ac|bd$, óf $ad|bc$ in T_2 . Door een tak van de ruggengraat van Q te verwijderen, hebben de twee deelbomen dus overlappende componenten in T_2 . Uit de definitie van een agreement forest volgt echter dat de bomen van $\mathcal{F}$ onderling knoop-disjunct zijn in T_2 . We hebben een tegenspraak.

We kunnen concluderen dat voor ieder kwartet Q in T_1 dat onverenigbaar is met T_2 is er tenminste één tak van A_Q verwijderd in T_1 .

Andersom, neem nu aan dat bepaalde takken van T_1 verwijderd zijn onder voorwaarden dat voor ieder kwartet Q in T_1 dat onverenigbaar is met T_2 is er tenminste één tak van A_Q verwijderd in T_1 . Laat $\mathcal{F} = \{t_1, \dots, t_k\}$ het bos zijn dat is ontstaan. We gaan laten zien dat $\mathcal{F}$ een agreement forest is. Dit doen we aan de hand van de definitie van een agreement forest.

1. Het is duidelijk dat $\mathcal{L}_1, \dots, \mathcal{L}_k \subseteq \mathcal{L}$.
2. We laten zien dat $t_j = T_1|_{\mathcal{L}_j} = T_2|_{\mathcal{L}_j}$ voor alle $j \in \{1, \dots, k\}$. Door de restrictie zitten er geen onverenigbare kwartetten van T_1 en T_2 in $\mathcal{F}$. Ieder kwartet in $\mathcal{F}$ is dus verenigbaar met T_2 , óf onsamenvhangend. Dus $\mathcal{F}$ en T_2 zijn per definitie verenigbaar.
3. We laten zien dat voor $i \in \{1, 2\}$ de bomen $\{T_1(\mathcal{L}_j) : j = 1, \dots, k\}$ knoop-disjunct zijn. We leiden tegenspraak af. Stel t_n, t_m in T_2 zijn niet tak-disjunct voor zekere $n, m \in \{1, \dots, l\}$ met $n \neq m$. Merk op dat t_n, t_m niet uit maar één blad kunnen bestaan, aangezien ze dan wel degelijk tak-disjunct zouden zijn. Kies nu bladeren $i, j \in t_n$ en $k, l \in t_m$ zó dat $Pad(i, j) \cap Pad(k, l) \neq \emptyset$ in T_2 . Omdat $i, j \in t_n$ en $k, l \in t_m$, zijn de takken wél disjunct in T_1 . Dit betekent dat $ijkl$ een onverenigbaar kwartet is. We hebben echter van ieder onverenigbaar kwartet tenminste één arm-tak verwijderd. Dit betekent dat $i, j \notin t_n$, óf $k, l \notin t_m$. We hebben een tegenspraak, dus t_n, t_m zijn tak-disjunct in T_2 . Omdat T_2 binair is, volgt dat t_n, t_m knoop-disjunct zijn.

We concluderen dat $\mathcal{F}$ een agreement forest is. □

5.2. ILP formulering

Lemma 5 impliceert dat een MAF gevonden kan worden door $|\mathcal{E}|$ te minimaliseren onder voorwaarde dat voor ieder kwartet Q in T_1 dat onverenigbaar is met T_2 er tenminste één tak van A_Q verwijderd is in T_1 . We introduceren voor elke $e \in E(T_1)$ een binaire variabele x_e met $x_e = 1$ als tak e verwijderd wordt en anders $x_e = 0$. De ILP formulering wordt dan:

$$\begin{aligned} & \min \sum_{i=1}^n x_i \\ & \text{o.d.v.:} \\ & \sum_{e_i \in A_{Q_j}} x_i \geq 1 \text{ voor alle onverenigbare kwartetten } Q_j; \\ & x_i \in \{0, 1\}. \end{aligned}$$

Aan de hand van deze ILP formulering kan een MAF gevonden worden van twee fylogenetische bomen. Zoals eerder gezegd geldt $d_{TBR}(T_1, T_2) = MAF(T_1, T_2)$. Het TBR-afstand probleem is dus opgelost.

Conclusie

De TBR-afstand tussen twee ongewortelde fylogenetische bomen T_1 en T_2 met dezelfde blad-verzameling is te bepalen aan de hand van een MAF. Een MAF is een agreement forest $\mathcal{F}$ bestaande uit zo min mogelijk deelbomen. Om de TBR-afstand te weten hoeft er slechts gekeken te worden naar het aantal componenten in een MAF.

Een MAF is te vinden door te kijken naar onverenigbare kwartetten. In een agreement forest $\mathcal{F}$ zitten namelijk slechts kwartetten die verenigbaar zijn met T_1 en T_2 . Er zijn twee methoden besproken om aan de hand van onverenigbare kwartetten een MAF te vinden. De eerste methode gebruikt Minimaal Onverenigbare Kwartetten. Dit zijn onverenigbare kwartetten met een langst mogelijke interne pad. Voor deze methode is een tweede fase nodig om een MAF te vinden. De fasen zitten ingewikkeld in elkaar en zijn lastig uit te voeren. Bovendien is het algoritme behorende bij hoofdstuk 4 een zogenaamd FPT algoritme, dat snel is wanneer de TBR afstand klein is, maar langzaam wanneer deze afstand groter is.

Het blijkt dat er een simpele methode is, welke in hoofdstuk 5 beschreven wordt. Aangezien onverenigbare kwartetten niet in een agreement forest zitten, moet er tenminste één tak verwijderd worden in een onverenigbaar kwartet. Minimaliseren we nu het aantal te verwijderen takken, hebben we een MAF gevonden. In deze methode wordt geheeltallig programmeren gebruikt. Voor een grote TBR afstand is de laatste methode simpeler en sneller dan de methode in hoofdstuk 4.

Referenties

1. Benjamin L. Allen and Mike Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of combinatorics*, 5(1): 1-15 (2001).
2. Yufeng Wu. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25(2): 190-196 (2009).
3. Michael Hallett and Catherine McCartin. A faster FPT algorithm for the maximum agreement forest problem. *Theory of Computing Systems* 41(3): 539-550 (2007).