Leo van Iersel
Faculteit Elektrotechniek, Wiskunde en Informatica
TU Delft
l.j.j.v.iersel@gmail.com

Hoe zijn ze verwant?

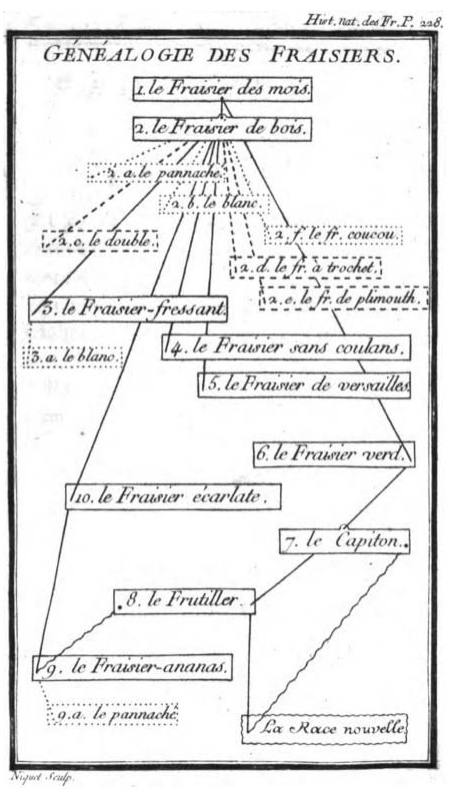
Hoe kunnen we reconstrueren hoe hedendaagse organismen zijn ontstaan uit verre voorouders door verschillende evolutionaire processen? Dit is het doel van onderzoek op het gebied van *fylogenetische netwerken*: grafen die evolutionaire verwantschappen beschrijven. Leo van Iersel ontving in 2011 een Veni-beurs van NWO voor onderzoek op dit gebied. Eind 2014 is hij begonnen als tenure track universitair docent aan de TU Delft. In dit artikel legt hij uit wat fylogenetische netwerken precies zijn en welke wiskundige problemen in dit vakgebied naar voren komen.

Darwins beroemde boek *The Origin of Species* (1859) bevatte precies één illustratie: een *fylogenetische boom*, oftewel een diagram dat schematisch weergeeft hoe soorten splitsen en zo nieuwe soorten vormen. Dit was niet de eerste keer dat zo'n diagram getekend werd: al voor de publicatie van *The Origin of Species* gaven biologen evolutionaire relaties weer in diagrammen. Interessant is dat die diagrammen lang niet altijd bomen waren.

In 1755 publiceerde Buffon al een diagram dat de evolutionaire relaties tussen verschillende hondenrassen weergaf [12]. Dit diagram was echter geen boom maar een netwerk. Hondenrassen splitsen namelijk niet alleen, maar nieuwe rassen kunnen ook gevormd worden uit combinaties van andere rassen. Een dergelijk diagram noemen we nu een *fylogenetisch netwerk*. De term 'fylogenetisch' bestaat uit de oud Griekse woorden *phulè* (volksstam) en *genesis* (wording). Een fylogenetisch netwerk beschrijft dus hoe groepen organismen (bijvoorbeeld stammen) zijn ontstaan uit andere groepen.

Deze netwerken zijn echter niet alleen relevant voor het bestuderen van verschillende stammen of rassen binnen één soort. Ook nieuwe soorten kunnen gevormd worden uit combinaties van andere soorten. Een mooi voorbeeld daarvan is de vorming van hybrides bij planten, zoals in het fylogenetische netwerk voor aardbeien in Figuur 1.

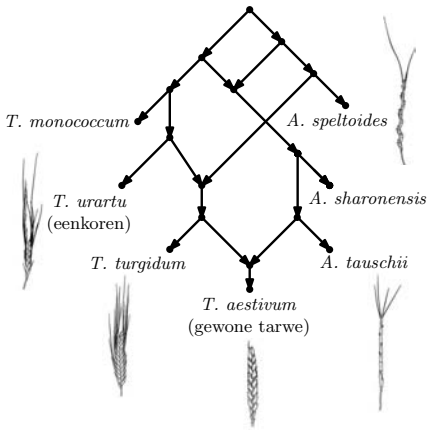
Toen Buffon honden bestudeerde, Duchenne aardbeien en Darwin vogels op de Galapagoseilanden, stelden ze allemaal dezelfde vraag: "Hoe zijn ze verwant?" Om dit uit te vinden moeten we achterhalen wat er miljoenen jaren geleden gebeurd is. Geen makkelijke taak. Nu, 260 jaar nadat Buffon zijn eerste fylogenetische netwerk publiceerde, zijn wiskundigen en informatici bezig om methodes te ontwikkelen om deze vraag systematisch te beantwoorden met behulp van DNA-data. Er is al veel bekend over het geval dat de verwantschappen beschreven kunnen worden door een fylogenetische boom. Het reconstrueren van fylogenetische netwerken is echter veel uitdagender. Niet alleen zijn er veel meer netwerken dan bomen, zelfs voor een gegeven netwerk is het vaak moeilijk om te bepalen hoe goed de data op dit netwerk passen. Pas recent zijn de eerste fylogenetische netwerken gepubliceerd die met behulp van computerprogramma's gegenereerd zijn, zoals het netwerk voor tarwe in Figuur 2.



Figuur 1 Een fylogenetisch netwerk getekend in 1766 door Antoine Nicolas Duchesne, dat de evolutionaire relaties beschrijft tussen verschillende aardbeiensoorten (in de ononderbroken rechthoeken) en -rassen [3, 14]. Het ras dat 'La Race nouvelle' wordt genoemd is door Duchesne op 17-jarige leeftijd ontdekt.

Wat is een fylogenetisch netwerk precies?

Er zijn verschillende definities van fylogenetische netwerken in omloop. Het belangrijkste onderscheid is dat tussen gerichte en ongerichte netwerken. Waar eerdere studies zich vooral concentreerden op ongerichte netwer-



Figuur 2 Een fylogenetisch netwerk dat laat zien hoe moderne tarwe is ontstaan uit een combinatie van verschillende oude tarwesoorten [13, 15]. De punten met twee inkomende pijlen beschrijven hybridisaties.

ken, wordt er steeds meer onderzoek gedaan over gerichte netwerken. De richtingen van de pijlen in zo'n netwerk geven de richting van evolutie aan. Ze geven dus een expliciete hypothese over de evolutionaire geschiedenis van de bestudeerde objecten. In dit artikel zal ik me beperken tot gerichte netwerken, die als volgt gedefiniëerd kunnen worden.

Definitie. Een *fylogenetisch netwerk* voor een verzameling X is een gerichte graaf met de volgende eigenschappen:

1. er zijn geen gerichte circuits;
2. er zijn geen punten met één inkomende en één uitgaande pijl (*overbodige* punten);
3. er is één punt zonder inkomende pijlen (de *wortel*);
4. de punten zonder uitgaande pijlen (de *bladeren*) zijn elk gelabeld met een element van X;
5. elk element van X is het label van één blad.

Een fylogenetisch netwerk is *binair* als elk punt hoogstens twee inkomende en hoogstens twee uitgaande pijlen heeft en elk punt met twee inkomende pijlen precies één uitgaande pijl heeft. Het netwerk voor tarwe uit Figuur 2 is een voorbeeld van een binair fylogenetisch netwerk.

In toepassingen in de biologie bevat de verzameling X een aantal namen van hedendaagse soorten, rassen of stammen. Voor het gemak gaan we uit van soorten. Elk blad stelt dus een hedendaagse soort voor. Een punt met meerdere uitgaande pijlen stelt een splitsing van een soort in twee of meer nieuwe soorten voor. Een punt met meer dan één inkomende pijl stelt voor dat een nieuwe soort ontstaan is uit een combinatie van eerdere soorten. Veelvoorkomende voorbeelden hiervan zijn de vorming van hybrides bij planten, reassortment bij virussen en genoverdracht bij bacteriën. Zo'n punt met minstens twee inkomende pijlen wordt een *reticulatie* genoemd.

Fylogenetische netwerken zijn een generalisatie van *fylogenetische bomen*, die we nu eenvoudig kunnen definiëren als fylogenetische netwerken zonder reticulaties. Fylogenetische bomen beschrijven net als fylogenetische netwerken evolutionaire relaties. Bomen zijn daarin echter veel beperkter dan netwerken. In een boom stelt elk intern punt een splitsing van een soort in twee of meer soorten voor. Het ontstaan van een soort uit een combinatie van eerdere soorten kan dus niet door een boom beschreven worden.

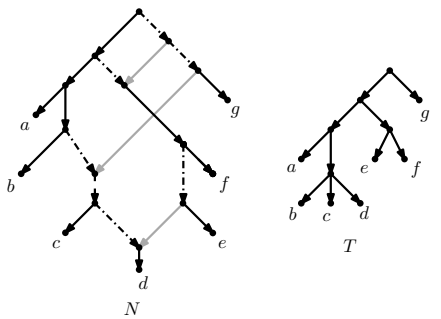
Bomen combineren tot een netwerk

Stel dat de evolutie van een verzameling X van soorten het best beschreven kan worden door een fylogenetisch netwerk N. Het DNA van deze soorten bestaat uit verschillende genen. Voor het gemak ga ik ervan uit dat genen in hun geheel worden overgeërfd. Als dat niet het geval is dan kunnen kleinere stukjes DNA beschouwd worden die wel in hun geheel worden overgeërfd.

In een punt van N met twee inkomende pijlen wordt dus een deel van de genen via de ene pijl geërfd en de rest van de genen via de andere pijl. De evolutie van een enkel gen kan daardoor beschreven worden door een boom, sterker nog, een boom die in het netwerk zit. Wat ik bedoel met 'in het netwerk' wordt geformaliseerd door de onderstaande definitie en geïllustreeerd in Figuur 3.

Definitie. Laat T een fylogenetische boom voor X zijn en N een fylogenetisch netwerk voor X. Dan is T *bevat* in N als we T kunnen verkrijgen uit N door het verwijderen van punten en pijlen en het samentrekken van pijlen.

Het *samentrekken* van een pijl van u naar v betekent dat je de pijlen die uit v vertrekken nu uit punt u laat vertrekken, de punten die in v aankomen nu in punt u laat aankomen, en vervolgens het punt v verwijdert. Waarom kan het nodig zijn om pijlen samen te trekken? Ten eerste, bij het verwijderen van punten en pijlen ontstaan er overbodige punten met één inkomende en één uitgaande pijl. Het *onderdrukken* van een overbodig punt betekent dat de inkomende pijl samengetrokken wordt. Alle overbodige punten moeten onderdrukt worden omdat deze niet toegestaan zijn in een fylogenetisch netwerk, en dus ook niet in een fylogenetische boom.



Figuur 3 Het fylogenetische netwerk N voor tarwe, met de bladeren voor het gemak herlabeld tot a,b,c,d,e,f en g en een fylogenetische boom T die bevat is in N. Boom T is bevat in N omdat je T uit N kunt verkrijgen door de grijze pijlen te verwijderen en alle onderbroken pijlen samen te trekken.

Er is echter nog een belangrijke reden om pijlen samen te trekken, wanneer de boom T niet binair is. Niet-binaire bomen worden in de praktijk gebruikt om onzekerheid uit te drukken. Bijvoorbeeld boom T in Figuur 3 geeft aan dat het onduidelijk is in welke volgorde b, c en d zijn afgesplitst van hun gemeenschappelijke voorouder. Dus in een netwerk dat deze boom bevat kunnen a, b en c in een willekeurige volgorde afsplitsen. Om T dan uit een deelgraaf van N te verkrijgen moeten pijlen samengetrokken worden.

Voor een gegeven netwerk en een gegeven boom is het trouwens al NP-moeilijk om te beslissen of het netwerk de boom bevat.

Een veel bestudeerde aanpak voor het construeren van fylogenetische netwerken uit DNA is de volgende tweestapsmethode. In de eerste stap wordt voor elk gen een fylogenetische boom gegenereerd. Voor deze stap zijn goede en snelle methodes beschikbaar. De tweede stap is om de verkregen fylogenetische bomen te combineren tot een fylogenetisch netwerk. Het doel is om een zo simpel mogelijk netwerk te vinden dat alle bomen bevat. Dit kunnen we als volgt formaliseren. Voor het gemak beperken we ons tot binaire netwerken.

Probleem: Minimum Reticulatie (MinRet).
Gegeven: een verzameling $\mathcal{T}$ van fylogenetische bomen, elk voor dezelfde verzameling X van labels.
Vind: een binair fylogenetisch netwerk N voor X dat elke boom in $\mathcal{T}$ bevat en een minimum aantal reticulaties heeft.

MinRet is een NP-moeilijk probleem, zelfs als $\mathcal{T}$ slechts twee binaire bomen bevat. Voor dit speciale geval is er echter een elegante karakterisering van het probleem, die gebruikt kan worden om het probleem relatief snel op te lossen.

Bos van overeenstemming

Stel $\mathcal{T}$ bestaat uit twee binaire bomen T_1 en T_2. Het idee is om pijlen uit T_1 en T_2 te verwijderen zodanig dat beide bomen in hetzelfde bos veranderen. Zo'n bos heet een 'bos van overeenstemming' omdat elke component van het bos in zekere zin consistent is met beide bomen. Beide bomen zijn het dus 'eens' over de evolutie van de soorten in één component van het bos.

Als T_1 en T_2 twee binaire bomen voor X zijn, dan is een *bos van overeenstemming* voor T_1 en T_2 een bos dat uit elk van T_1 en T_2 verkregen kan worden door een deel van de pijlen en ongelabelde punten te verwijderen en overbodige punten te onderdrukken.

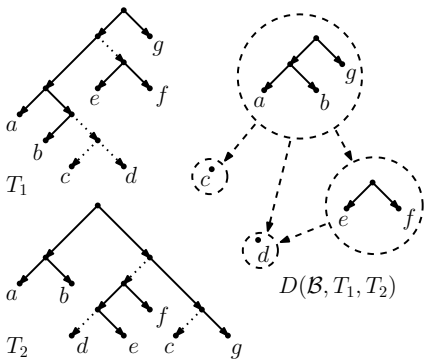
Stel nu dat N een netwerk is dat T_1 en T_2 bevat. Stel dat we voor elk punt in N met twee inkomende pijlen beide pijlen verwijderen, en vervolgens alle ongelabelde bladeren verwijderen en overbodige punten onderdrukken totdat er geen ongelabelde bladeren of overbodige punten meer zijn. Dan wordt N veranderd in een bos. Sterker nog, we verkrijgen een bos van overeenstemming voor T_1 en T_2.

Het is nu verleidelijk om te denken dat we uit elk bos van overeenstemming voor T_1 en T_2 ook een netwerk kunnen maken dat beide bomen bevat. Dit is echter alleen mogelijk als het bos aan de volgende acycliciteitsvoorwaarde voldoet. We definiëren een gerichte graaf $D(\mathcal{B}, T_1, T_2)$ die beschrijft hoe de componenten van $\mathcal{B}$ zich verhouden in de bomen T_1 en T_2. Deze gerichte graaf $D(\mathcal{B}, T_1, T_2)$ heeft een punt voor elke component van $\mathcal{B}$ en een pijl van (het punt voor) een component C_1 naar (het punt voor) een component C_2 als er een gericht pad is van de wortel van C_1 naar de wortel van C_2 in tenminste één van T_1 en T_2.

Definitie. Een bos van overeenstemming $\mathcal{B}$ voor T_1 en T_2 is *acyclisch* als de gerichte graaf $D(\mathcal{B}, T_1, T_2)$ acyclisch is (dat wil zeggen geen gerichte circuits bevat).

Stelling (Baroni e.a. [2]). *Als T_1 en T_2 twee binaire bomen voor X zijn, dan bestaat er een binair netwerk met k reticulaties dat T_1 en T_2 bevat dan en slechts dan als T_1 en T_2 een acyclisch bos van overeenstemming hebben met $k+1$ componenten.*

Het oplossen van MinRet is dus equivalent aan het vinden van een acyclisch bos van overeenstemming met zo min mogelijk componenten. Zo'n bos noemen we een *acyclisch bos van maximum overeenstemming*.



Figuur 4 Deze twee fylogenetische bomen T_1 en T_2 hebben een bos van overeenstemming $\mathcal{B}$ met vier componenten. Dit zijn de componenten die je krijgt als je de gestippelde lijnen uit T_1 en T_2 verwijdert en vervolgens overbodige punten onderdrukt. De gerichte graaf $D(\mathcal{B},T_1,T_2)$ geeft aan hoe de componenten van $\mathcal{B}$ zich verhouden in T_1 en T_2. Omdat $D(\mathcal{B},T_1,T_2)$ geen gerichte circuits bevat is $\mathcal{B}$ een *acyclisch* bos van overeenstemming. Volgens de stelling van Baroni e.a. is er dus een netwerk met drie reticulaties dat T_1 en T_2 bevat (namelijk het netwerk N uit Figuur 3).

De acycliciteit blijkt het belangrijkste obstakel te zijn voor het vinden van een efficiënt benaderingsalgoritme voor MinRet. Dit probleem blijkt namelijk net zo moeilijk te benaderen als het probleem Directed Feedback Vertex Set (DFVS): maak een gerichte graaf acyclisch door zo min mogelijk punten te verwijderen.

Stelling [11]. *Er bestaat een constante-factor-benaderingsalgoritme voor* MinRet *beperkt tot twee binaire bomen dan en slechts dan als een dergelijk algoritme bestaat voor* DFVS.

Een algoritme is een *constante-factor benaderingsalgoritme* als er een constante c bestaat zodanig dat het algoritme in polynomiale tijd een oplossing vindt die maximaal c maal slechter is dan een optimale oplossing. Of voor MinRet en DFVS een dergelijk algoritme bestaat is een belangrijk open probleem. DFVS was één van de eerste 21 problemen waarvan is bewezen dat ze NP-volledig zijn, door Richard Karp in een beroemd artikel uit 1971 [10], maar nog steeds is het niet bekend of er een constante-factor-benaderingsalgoritme voor bestaat.

Gelukkig is DFVS in de praktijk goed op te lossen met behulp van geheeltallig programmeren. In combinatie met een benaderingsalgoritme voor het vinden van een bos van maximum overeenstemming geeft dit een praktisch benaderingsalgoritme voor MinRet beperkt tot twee binaire bomen, wat zelfs uitgebreid kan worden voor niet-binaire bomen [7].

Door de bomen het bos niet meer zien

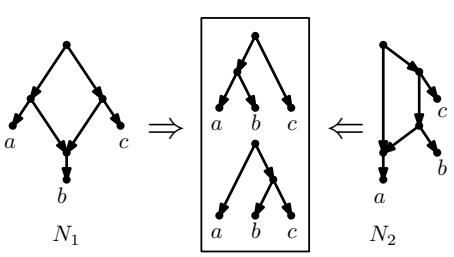
We hebben gezien dat voor het oplossen van MinRet de relatie met acyclische bossen van overeenstemming van groot belang is. Helaas bestaat deze relatie, beschreven in de stelling van Baroni e.a., alleen voor instanties van twee bomen. Voor instanties met drie of meer bomen geldt de relatie nog maar één kant uit. Het aantal componenten in een acyclisch bos van maximum overeenstemming geeft alleen een ondergrens op het aantal reticulaties in een optimaal netwerk. Zelfs voor instanties bestaande uit drie binaire bomen wordt MinRet erg uitdagend. Het theoretisch snelste algoritme voor dit geval heeft looptijd $1609891840^k p(n)$, met k het aantal reticulaties in een optimaal netwerk en $p(n)$ een polynoom in het aantal bladeren n [8]. Voor algemene instanties, waarin een willekeurig aantal niet-binaire bomen is toegestaan, is het niet bekend of er een algoritme met looptijd $f(k)\cdot p(n)$ bestaat met f een functie van k en p een polynoom in n.

Welke informatie is nodig?

Als we willen reconstrueren hoe de evolutie van een groep soorten er precies uit heeft gezien, dan maken we een fylogenetisch netwerk. Maar hoe weten we zeker dat we het juiste netwerk hebben? In welke gevallen wordt een netwerk uniek bepaald door de data? Als de data, zoals hierboven, bestaan uit fylogenetische bomen, dan kan het zijn dat het netwerk uniek bepaald is maar in het simpele voorbeeld in Figuur 5 is dat bijvoorbeeld niet zo.

Voor fylogenetische bomen is er veel onderzoek gedaan naar dit soort vraagstukken. Een fylogenetische boom wordt bijvoorbeeld uniek bepaald door de verzameling *triplets* die het bevat. *Triplets* zijn fylogenetische bomen met elk drie bladeren. Bovendien is er een polynomiale-tijd-algoritme (Aho e.a. [1]) om, gegeven een willekeurige verzameling triplets, te bepalen of er een fylogenetische boom bestaat die deze triplets bevat.

Dit wordt gebruikt voor zogenaamde 'superboom'-methodes. Stel dat voor een aantal verschillende deelverzamelingen van X een fylogenetische boom bekend is. Kunnen de-



Figuur 5 Twee fylogenetische netwerken N_1 en N_2 voor $\{a,b,c\}$ die allebei precies dezelfde verzameling bomen bevatten.

ze fylogenetische bomen dan samengevoegd worden tot een fylogenetische 'superboom' voor X die elke gegeven boom bevat? Deze vraag kunnen we nu gemakkelijk beantwoorden. Eerst vinden we voor elke invoerboom de verzameling triplets die het bevat. Daarna bepalen we of er een fylogenetische boom bestaat die de vereniging van de verkregen triplet verzamelingen bevat, met het algoritme van Aho e.a. Voor het geval dat er geen boom bestaat die *alle* triplets bevat zijn er tal van heuristieken ontwikkeld die toch een redelijke superboom genereren.

Maar waardoor wordt een fylogenetisch netwerk uniek bepaald? En kunnen we 'supernetwerk'-methodes ontwikkelen?

Elk fylogenetisch netwerk voor X induceert, voor elke deelverzameling X' van X, een fylogenetisch netwerk voor X', volgens de volgende definitie. Laat $LSV(X')$ (Laatste Stabiele Voorouder) het laatste punt zijn dat ligt op alle gerichte paden van de wortel van het netwerk naar een blad met label in X'.

Definitie. Gegeven een fylogenetisch netwerk N voor X en een deelverzameling $X' \subsetneq X$, wordt het *deelnet* $N|X'$ verkregen door

1. alle punten en pijlen te nemen die op een gericht pad liggen van $LSV(X')$ naar een blad met label in X';
2. alle overbodige punten te onderdrukken en parallelle pijlen te vervangen door enkele pijlen totdat er geen overbodige punten of parallelle pijlen meer zijn.

De verzameling deelnetten geïnduceerd door een netwerk N voor X is nu gedefinieerd als

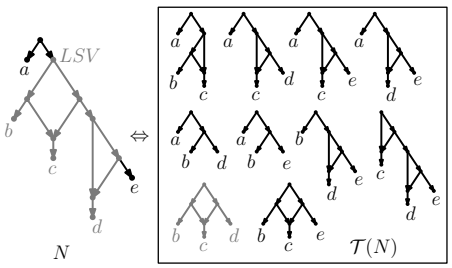
$$\mathcal{S}(N) = \{N|X' : X' \subsetneq X\}.$$

Net zo als triplets fylogenetische bomen zijn met drie bladeren, kunnen we *trinetten* definiëren als fylogenetische netwerken met drie bladeren. De verzameling trinetten geïnduceerd door een fylogenetisch netwerk N wordt dan gedefinieerd als

$$\mathcal{T}(N) = \{N|X' : X' \subsetneq X, |X'| = 3\}.$$

Helaas blijkt dat fylogenetische netwerken in het algemeen niet uniek bepaald worden door de verzameling trinetten die ze induceren, en ook niet door de hele verzameling geïnduceerde deelnetten [6].

De tegenvoorbeelden zijn echter erg complex: het aantal reticulaties groeit exponentieel in het aantal bladeren. Redelijk simpele netwerken worden wel uniek bepaald



Figuur 6 Een fylogenetisch netwerk N en de verzameling $\mathcal{T}(N)$ van alle trinetten die het bevat. Laat bijvoorbeeld $X'=\{b,c,d\}$. Dan is het punt LSV het laatste punt dat ligt op alle gerichte paden van de wortel van N naar een blad met label in $\{b,c,d\}$. Alle punten en pijlen die op een gericht pad liggen van LSV naar een blad met label in $\{b,c,d\}$ zijn grijs gekleurd. Als we in deze grijze deelgraaf nu alle overbodige punten onderdrukken en parallelle pijlen vervangen door enkele pijlen totdat er geen overbodige punten of parallelle pijlen meer zijn, dan verkrijgen we het grijze deelnet $N|\{b,c,d\} \in \mathcal{T}(N)$. Het blijkt dat in dit geval N uniek bepaald wordt door $\mathcal{T}(N)$.

door $\mathcal{T}(N)$ [9]. Dit is bijvoorbeeld het geval voor netwerken met maximaal twee reticulaties per 2-samenhangende deelgraaf (*2-samenhangend* betekent dat de deelgraaf samenhangend blijft als je een willekeurige pijl verwijdert). Het geldt ook voor netwerken waarin elk punt dat geen blad is tenminste één uitgaande pijl heeft naar een punt dat geen reticulatie is. Bijvoorbeeld de verzameling $\mathcal{T}(N)$ van trinetten in Figuur 6 bepalen het netwerk N in de figuur, want N voldoet aan beide voorwaarden. Het is echter onbekend waar precies de grens ligt tussen netwerken die wel en niet uniek bepaald worden door hun verzameling geïnduceerde trinetten.

Dit soort vraagstukken zijn van groot belang voor toepassingen in de biologie. Een belangrijke eis aan een methode voor het maken van fylogenetische netwerken (of bomen) is dat de methode *consistent* is, dat wil zeggen dat de methode het juiste netwerk produceert indien het volledige en foutloze data krijgt aangeboden. In deze zin kan een methode die fylogenetische bomen combineert tot een fylogenetisch netwerk nooit consistent zijn. Hetzelfde geldt voor een 'supernetwerk' methode die trinetten of deelnetwerken combineert tot een volledig netwerk. Maar als we ons beperken tot de genoemde klassen van netwerken die wel uniek bepaald worden door hun trinetten, dan is het wel mogelijk om consistente supernetwerk-methodes te ontwikkelen.

Kunnen we het DNA rechtstreeks gebruiken?

Voor het maken van fylogenetische bomen zijn tal van methodes ontwikkeld. Tegenwoordig is de belangrijkste bron van data natuurlijk DNA. Er zijn methodes die in twee stappen met het DNA werken. In een eerste stap kan bijvoorbeeld een afstand tussen elk tweetal soorten berekend worden, of een triplet voor elk drietal soorten, op basis van het DNA. Een tweede stap is dan om een boom te vinden die zo goed mogelijk voldoet aan de afstanden of triplets. De meest nauwkeurige methodes werken echter rechtstreeks met het DNA. Ze zoeken een boom die een zekere score maximaliseert. De score is een functie van de boom en de DNA-sequenties, en kan bijvoorbeeld gebaseerd zijn op een waarschijnlijkheidsberekening of op het *parsimony*-principe.

Het idee van *parsimony* (gierigheid) is om een fylogenetische boom te vinden waarop het gegeven DNA met zo min mogelijk mutaties geëvolueerd zou kunnen zijn. Het grote voordeel van parsimony is dat de parsimonyscore van een fylogenetische boom en gegeven DNA-sequenties in polynomiale tijd berekend kan worden. Merk eerst op dat de posities in de DNA-sequenties onderling onafhankelijk zijn en dat je dus elke positie apart kunt bekijken. Voor elke positie moet je dan het volgende probleem oplossen. Hierin is $\mathcal{P}$ de verzameling mogelijke letters in de sequenties. Dus bijvoorbeeld voor DNA is $\mathcal{P} = \{A, C, G, T\}$.

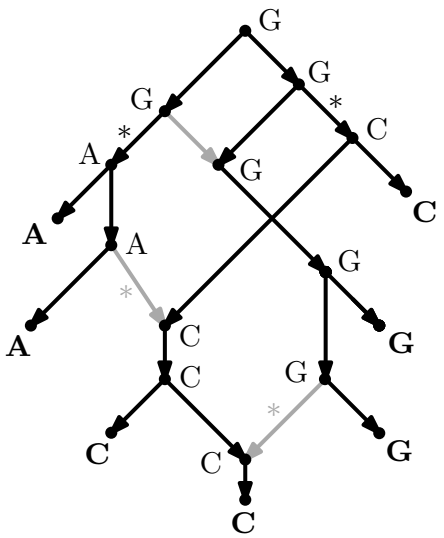
Probleem: Parsimony voor Bomen.
Gegeven: een fylogenetische boom en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een label $\ell(v) \in \mathcal{P}$ voor elk intern punt v zodanig dat het aantal pijlen (u, w) met $\ell(u) \neq \ell(w)$ minimaal is.

Dit minimum aantal wordt de *parsimonyscore* genoemd. Dit probleem kan in polynomiale tijd opgelost worden door middel van dynamisch programmeren [5].

Voor fylogenetische netwerken zijn er twee interessante generalisaties van deze score. De eerste generalisatie is wiskundig gezien misschien de meest logische.

Probleem: Netwerk Parsimony.
Gegeven: een fylogenetisch netwerk en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een label $\ell(v) \in \mathcal{P}$ voor elk intern punt v zodanig dat het aantal pijlen (u, w) met $\ell(u) \neq \ell(w)$ minimaal is.

Als $|\mathcal{P}| = 2$ dan is dit probleem direct gerelateerd aan het bekende MinCut-probleem, waarin een ongerichte graaf gegeven is met twee speciale punten s en t en gevraagd wordt om zo min mogelijk lijnen uit de graaf te verwijderen zodat er geen pad tussen s en t meer bestaat. Hoe is dit gerelateerd aan Netwerk Parsimony? Stel dat een netwerk is ge-



Figuur 7 Stel het fylogenetische netwerk voor tarwe is gegeven samen met de (vetgedrukte) labels van de bladeren. De figuur geeft één mogelijke labelling van de interne punten. Voor deze labelling zien we dat er in totaal vier pijlen zijn waar het label verandert, gemarkeerd met een $*$. De optimale waarde van Netwerk Parsimony is dus maximaal 4. Als we echter naar de boom kijken die bestaat uit de zwarte pijlen, dan zijn er maar twee pijlen waar het label verandert (de grijze $*$'en). De optimale waarde van Boom-in-netwerk Parsimony is dus maximaal 2.

geven met een label uit $\mathcal{P}$ voor elk blad. Neem voor het gemak aan dat $\mathcal{P} = \{S, T\}$. Stel nu dat we alle bladeren met label S samenvoegen tot een enkel punt dat we s noemen, en alle bladeren met label T samenvoegen tot een enkel punt dat we t noemen. Dan komt het oplossen van MinCut in de verkregen graaf op hetzelfde neer als Netwerk Parsimony oplossen in het oorspronkelijke netwerk [4].

Voor $|\mathcal{P}| = 2$ is dit probleem dus in polynomiale tijd op te lossen door middel van het bekende Edmonds–Karp-algoritme voor MaxFlow. Voor $|\mathcal{P}| > 2$ wordt het probleem NP-moeilijk maar is het nog wel goed benaderbaar.

Vanuit biologisch perspectief is de volgende generalisatie echter logischer.

Probleem: Boom-in-netwerk Parsimony.
Gegeven: een fylogenetisch netwerk N voor X en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een fylogenetische boom voor X die bevat is in N en een zo klein mogelijke parsimony score heeft.

Waarom is Boom-in-netwerk Parsimony biologisch gezien een logischere generalisatie? Kijk wat er bij de reticulaties gebeurd met de sequenties. Hier worden de sequenties van twee 'voorouders' gecombineerd tot de sequentie van een 'hybride'. Kijk je echter naar een enkele positie van het DNA, dan wordt deze van een van de twee voorouders geërfd. De evolutie van een enkele DNApositie kan dus altijd beschreven worden door een boom — een boom die bevat is in het netwerk. Helaas is dit Boom-in-netwerk Parsimony probleem veel moeilijker op te lossen of te benaderen dan Netwerk Parsimony [4]. Alleen als we ons beperken tot netwerken met weinig reticulaties per 2-samenhangende deelgraaf dan kunnen we deze score snel berekenen.

Tot slot

Het reconstrueren van fylogenetische netwerken blijkt veel moeilijker te zijn dan het maken van fylogenetische bomen. Netwerken zijn tenslotte veel complexer dan bomen. Toch kunnen we in veel gevallen ver komen zolang we ons beperken tot relatief simpele netwerken. Dit is nuttig in de biologie omdat veel praktische fylogenetische netwerken vrij simpel zijn, zoals bijvoorbeeld de netwerken voor aardbeien en tarwe uit het begin van dit artikel. Toch komen er ook veel complexere fylogenetische netwerken voor in de biologie, bijvoorbeeld voor bacteriën. Zulke netwerken zullen we waarschijnlijk nooit tot in alle details kunnen reconstrueren. Dit is echter ook niet altijd noodzakelijk voor biologen. Belangrijker is het om het netwerk in hoofdlijnen te kunnen schetsen, zodat biologen er belangrijke conclusies uit kunnen trekken over de relaties tussen verschillende soorten.

Referenties

1 A.V. Aho, Y. Sagiv, T.G. Szymanski en J.D. Ullman, Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, *SIAM Journal on Computing* 100(3) (1981), 405–421.

2 M. Baroni, S. Grünewald, V. Moulton en C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, *Journal of Mathematical Biology* 51(2) (2005), 171–182.

3 A.N. Duchesne, *Histoire naturelle des fraisiers, contenant Les vues d'Économie réunies à la Botanique; et suivie de Remarques Particulières sur plusieurs points qui ont rapport à l'Histoire naturelle générale*, Didot le jeune & C.J. Panckoucke, Parijs, 1766.

4 M. Fischer, L.J.J. van Iersel, S.M. Kelk en C. Scornavacca, On Computing the Maximum Parsimony Score of a Phylogenetic Network, *SIAM Journal on Discrete Mathematics* 29(1) (2015), 559–585.

5 W. Fitch, Toward defining the course of evolution: Minimum change for a specific tree topology, *Systematic Zoology* 20 (1971), 406–416.

6 K.T. Huber, L.J.J. van Iersel, V. Moulton en T. Wu, How Much Information is Needed to Infer Reticulate Evolutionary Histories? *Systematic Biology* 64(1) (2015), 102–111.

7 L.J.J. van Iersel, S.M. Kelk, N. Lekić en C. Scornavacca, A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees, *BMC Bioinformatics* 15 (2014), 127.

8 L.J.J. van Iersel, S.M. Kelk, N. Lekić, C. Whidden en N. Zeh, Hybridization Number on Three Trees (2014), arXiv:1402.2136 [cs.DS].

9 L.J.J. van Iersel en V. Moulton, Trinets encode tree-child and level-2 phylogenetic networks, *Journal of Mathematical Biology* 68(7) (2014), 1707–1729.

10 R.M. Karp, Reducibility Among Combinatorial Problems, in R.E. Miller en J.W. Thatcher (eds.), *Complexity of Computer Computations*, Plenum, New York, 1972, pp. 85–103.

11 S.M. Kelk, L.J.J. van Iersel, N. Lekić, S. Linz, C. Scornavacca en L. Stougie, Cycle killer... qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set, *SIAM Journal on Discrete Mathematics* 26(4) (2012), 1635–1656.

12 G.-L. Leclerc, graaf de Buffon, *Histoire naturelle générale et particulière*, Vol. 5, Imprimerie Royale, Parijs, 1755, pp. 228–229.

13 T. Marcussen, S.R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, International Wheat Genome Sequencing Consortium, K.S. Jakobsen, B.B. Wulff, B. Steuernagel, K.F. Mayer en O.A. Olsen, Ancient hybridizations among the ancestral genomes of bread wheat, *Science* 345 (2014), 1250092.

14 D. Morrison, The second phylogenetic network (1766), phylonetworks.blogspot.nl/2012/04/second-phylogenetic-network-1766.html.

15 D. Morrison, Complex hybridizations in wheat, phylonetworks.blogspot.nl/2015/01/complex-hybridizations-in-wheat.html.