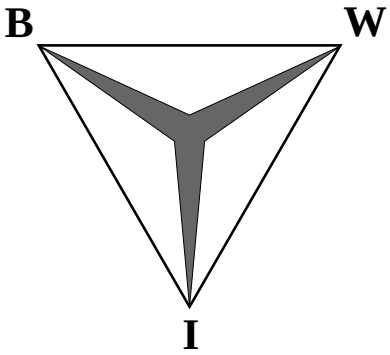
Oncogenetische Bomen

Een wiskundig model voor de ontwikkeling van kanker

Joost van Hooff

12 juli 2002



BWI-werkstuk
Vrije Universiteit
Faculteit der Exacte Wetenschappen
Divisie Wiskunde en Infomatica
Studierichting Bedrijfswiskunde & Infomatica
De Boelelaan 1081a
1081 HV Amsterdam

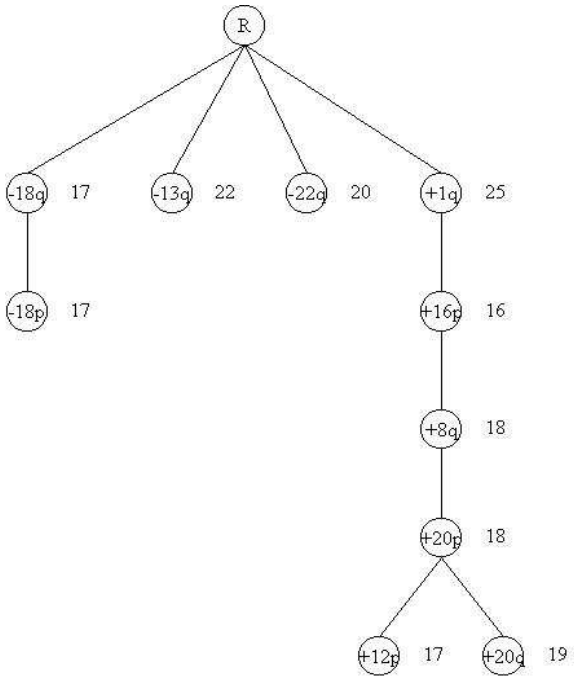
Voorwoord

Het BWI-werkstuk is één van de laatste onderdelen van de studie Bedrijfswiskunde & Informatica (BWI). Als het goed is wordt dit onderdeel het semester voor de afstudeerstage gedaan. Zo is het ook het geval bij mijzelf. Het doel van het werkstuk is dat de student voor een deskundige manager op heldere wijze een probleem beschrijft. Onder een deskundige manager wordt verstaan dat de manager beschikt over een algemene kennis over het onderwerp. Ik ben er daarbij vanuit gegaan dat zo'n deskundige manager over dezelfde kennis beschikt als dat ik zelf had over dit onderwerp, voordat ik aan het werkstuk begon. Dat betekent dat ik er vanuit ga dat de manager een goede wiskundige kennis beschikt en over een mindere of weggezakte kennis van biologie en genetica in het bijzonder. Verder houdt op heldere wijze in dat het werkstuk beknopt moet zijn. Aangezien de werkelijke tekst van dit werkstuk slechts 24 pagina's beslaat, lijkt mij dit aardig gelukt. Voor iemand die over helemaal weinig tijd beschikt, volgt na dit voorwoord eerst een samenvatting.
Doelstelling van dit BWI-werkstuk is het beschrijven van de methode CGH (Comparitive Genomic Hybridization) en het wiskundige model voor analyse van de CGH-data, en een toepassing van dit model op data van de VU. Ik heb dit onderwerp gekozen, omdat ik voordat ik BWI ging studeren erg getwijfeld heb tussen de studie Biologie en BWI. Het leek mij daarom leuk om datgene wat ik tijdens mijn studie heb geleerd toe te passen op een biologisch onderwerp. Het onderwerp heb ik niet zelf bedacht, maar stond op de website van het stagebureau.
Nu weet ik dat het bedrijfskundige gedeelte van mijn studie helemaal niet terug zal komen in dit werkstuk. Wel is dit werkstuk volgens mij een goede combinatie tussen wiskunde en informatica. Door de toepassing van de theorie op de praktijk lijkt mij dit werkstuk toch een echt BWI-werkstuk. Verder is het zo dat dit werkstuk dan wel geen bedrijfskundige kant behandeld, maar wel een maatschappelijke. Kanker is namelijk in ons land, en in de rest van de wereld, een belangrijke doodsoorzaak. Het is daarom te hopen dat er door onderzoek te doen, deze ziekte ooit goed te genezen valt. Nu zal mijn werkstuk daar geen bedrage aan leveren, maar het is wel een teken dat iemand die de studie BWI heeft gedaan hier wel zijn bijdrage aan zou kunnen leveren.
Als laatste wil ik nog twee mensen bedanken. Ten eerste natuurlijk mijn begeleider Elena Marchiori. Ik wil haar bedanken voor het kritisch doorlezen van mijn werkstuk, het helpen bij het uitpluizen van de (vele) wiskundige bewijzen en voor het feit dat zij mij in contact bracht met mensen die wat meer verstand hadden van sommige onderwerpen dan ons. Ook zou ik Kees Jong willen bedanken voor de uitleg die hij gegeven heeft over CGH en voor helpen bij het gebruik van CGH-data voor mijn onderzoek.
Ik wens iedereen veel plezier met het lezen van dit werkstuk,

Joost van Hooff, Juni 2002

Samenvatting

Kanker ontstaat doordat er zich in het DNA een verandering voordoet. Deze verandering gebeurt onder invloed van bepaalde stoffen of straling. Een verandering in het DNA houdt in dat er op een bepaalde plaats van het DNA een stukje DNA vermeerderd, een gain, of dat er juist een stukje DNA verloren gaat, een loss. Ligt er op een gain een kankerbevorderend gen of ligt er op een loss een kankeronderdrukkend gen, dan kan er kanker ontstaan. Heeft er zich eenmaal een gain of een loss in het DNA voorgedaan, dan wordt dit gevolgd door meerdere veranderingen in DNA.
Tien jaar geleden is er een nieuwe methode ontwikkeld om deze gains en losses in het DNA op te sporen, Comparative Genomic Hybridization (CGH). Door zowel gezond als tumor weefsel te laten hybridiseren met gezond weefsel, is men door middel van kleuring in staat om de gains en losses in DNA te ontdekken. CGH is, in tegenstelling tot andere methodes, simpeler te gebruiken en kan met deze methode al het genetisch materiaal van een cel in één keer worden onderzocht. Daarom heeft CGH al snel aan populariteit gewonnen en zijn er inmiddels al belangrijke ontdekkingen mee geboekt.
Men vermoedt dat de veranderingen in het DNA zich niet random voor doen, maar dat er een bepaald patroon in te ontdekken valt. Door de CGH-resultaten van meerdere tumoren te vergelijken, zou zo'n patroon ontdekt kunnen worden. Een model voor dat patroon zijn oncogenetische bomen. Hierbij is elk punt in de boom een gain of een loss in een bepaalde arm van het DNA. Van deze bomen bestaat zowel een tijdsonafhankelijke als een tijdsafhankelijke variant. Er kan bewezen worden dat als een tijdsafhankelijke oncogenetische boom een padstructuur heeft, dat er dan ook een tijdsonafhankelijke variant bestaat.
Er bestaat een algoritme om een niet schuine boom te generen uit CGH data. Onder een niet schuine boom wordt een boom verstaan waarbij de topologie van de boom overeenkomt met de kansverdeling van de boom. Bij dit algoritme worden eerst de kansen berekend dat een event, een gain of een loss, wordt geobserveerd. Ook de kans dat twee events samen worden geobserveerd worden berekend. Vervolgens worden, door middel van een gewicht functie, gewichten toegekend aan de lijnen tussen de events. Daarna wordt door middel van maximum branching de boom bepaald met het hoogste gewicht. Er kan bewezen worden dat dit algoritme zowel in theorie werkt. Ook er kan bewezen worden dat dit algoritme in praktijk, mits er voldoende samples beschikbaar zijn, een niet schuine boom met hoge kans juist construeert.
Met behulp van software is het mogelijk om een niet schuine boom uit CGH data te construeren. De gebruikte CGH-data is afkomstig van borstkanker. De beste boom die we hieruit vinden is de volgende



Door een aantal beperkingen, waaronder een tekort aan samples, kunnen we niet concluderen dat deze boom in zijn geheel juist is. Wel kunnen we concluderen dat de gain in 1q en de losses in 13q en 22q in het beginstadium van kanker plaatsvinden. Dit wordt gedeeltelijk ook bevestigd door eerder uitgevoerd onderzoek.

Inhoudsopgave

Hoofdstuk 1

Inleiding

Kanker ontstaat doordat er zich in het DNA een verandering voordoet. Deze verandering kan ontstaan als gevolg van bijvoorbeeld radioactieve straling. Door deze ene verandering kunnen er ook meerdere veranderingen ontstaan. Sinds tien jaar is er een nieuwe methode waarmee het mogelijk is om deze veranderingen in het DNA te vinden. Deze methode wordt comparative genomic hybridization (CGH) genoemd. Voor de resultaten van CGH is een wiskundig model opgesteld. Met dit model is het mogelijk om te achterhalen in welke volgorde zich de veranderingen in het DNA hebben voorgedaan. Hiervoor is een algoritme opgesteld. In dit BWI-werkstuk zullen we antwoord proberen te geven op de volgende drie vragen:

1. Wat houdt de methode CGH precies in?
2. Welk wiskundig model wordt er gebruikt voor het analyseren van de CGH-data?
3. Welke resultaten geeft dit model bij toepassing op data van de VU?

Doelstelling van dit BWI-werkstuk is dus het beschrijven van de methode CGH en het wiskundige model voor analyse van de CGH-data, en een toepassing van dit model op data van de VU.
In Hoofdstuk 2 beginnen we met het geven van achtergrondinformatie over genetisch materiaal van de mens en over kanker. In Paragraaf 2.1 zullen we kort uitleggen wat chromosomen en DNA precies zijn en wat hun functie zijn in het menselijk lichaam. Vervolgens zullen we in Paragraaf 2.2 kort uitleggen wat kanker precies is en welke oorzaken het heeft. Voor de lezer met een biologische achtergrond is het geen probleem als deze Hoofdstuk 2 overslaat.
In Hoofdstuk 3 wordt de methode CGH beschreven en worden de belangrijkste resultaten van deze methode besproken. Paragraaf 3.1 beschrijft één bepaalde CGH methode, er bestaan namelijk meerdere. De methode die in deze paragraaf beschreven wordt is array CGH. In Paragraaf 3.2 worden de belangrijkste resultaten van CGH beschreven. Ook worden een aantal verwachtingen besproken die men met CGH nog hoopt te kunnen behalen.
Vervolgens wordt in Hoofdstuk 4 het wiskundige model besproken. We beginnen in Paragraaf 4.1 met het definiëren van oncogenetische bomen. Deze bomen zijn namelijk de uitkomst van het later te definiëren algoritme. Vervolgens bewijzen we in Paragraaf 4.1.1 dat een tijdsafhankelijke oncogenetische boom met een padstructuur dezelfde kansverdeling genereerd als tijdsonafhankelijke boom met een padstructuur. In Paragraaf 4.2 komt het werkelijke reconstructie probleem aan de orde. Deze paragraaf is opgedeeld in twee delen. In het eerste deel, Paragraaf 4.2.1 bespreken we een aantal problemen die op kunnen treden bij de reconstructie. Vervolgens bespreken we enige aannames om deze problemen te voorkomen. In Paragraaf 4.2.2 komt dan het reconstructie algoritme aan bod. We bewijzen dat het algoritme voor een bepaalde gewichtfunctie in theorie werkt. Verder bewijzen we ook onder welke voorwaarde dit algoritme in de praktijk werkt. Voor de lezer zonder wiskundige achtergrond is het aan te raden om Paragraaf 4.1.1 en 4.2 over te slaan.
In Hoofdstuk 5 volgt dan de toepassing van het algoritme op CGH-data van de VU. De data is afkomstig van Bauke Ylstra en is vervolgens bewerkt door Kees Jong. De analyse is uitgevoerd met behulp van software die ik gekregen heb van de schrijvers van artikel [4]. De data die onderzocht

wordt is afkomstig van borstkanker tumoren. In Paragraaf 5.1 wordt een omschrijving van de data gegeven en van wat er uit de data is geselecteerd. Paragraaf 5.2 beschrijft de toepassing van het algoritme en de resultaten die daaruit verkregen werden. Tenslotte worden in Paragraaf 5.3 conclusies uit de resultaten getrokken. Ook zal daar kritisch bij het onderzoek worden stilgestaan.

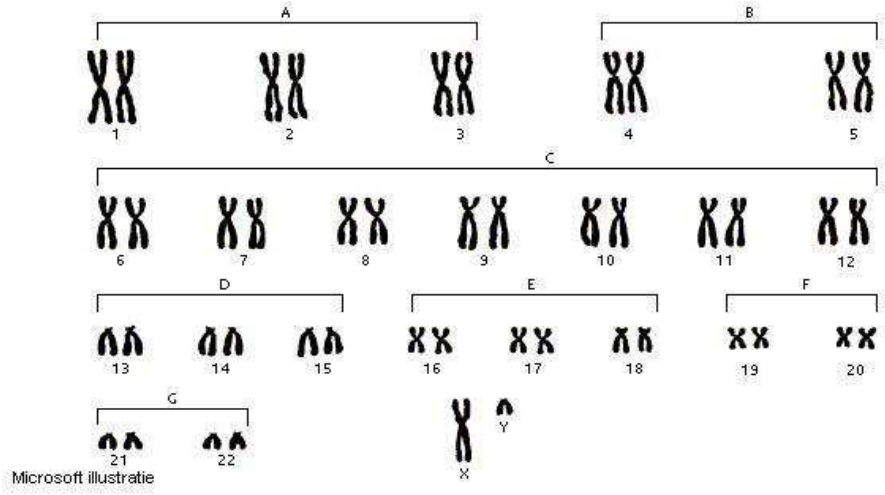
Hoofdstuk 2

Algemene Informatie

Voordat we beginnen met het beantwoorden van de onderzoeksvragen, gaan we eerst beginnen met het behandelen van wat algemene informatie. Bij een BWI-werkstuk wordt er namelijk vanuit gegaan dat men beschikt over een aardige kennis van de wiskunde, maar niet van de biologie en de genetica in het bijzonder. Daarom zullen we in Paragraaf 2.1 een aantal belangrijke genetische begrippen op een rijtje zetten. Vervolgens zullen we in Paragraaf 2.2 omschrijven wat kanker ongeveer inhoudt. Beide Paragrafen zijn gebaseerd op een combinatie van twee bronnen, namelijk [1] en [7].

2.1 Chromosomen en DNA

Al het erfelijk materiaal van een mens ligt opgeslagen in zijn chromosomen. Elke cel in het menselijk lichaam heeft 23 paren chromosomen. Hiervan zijn 22 paren identiek zijn. Het laatste paar kan verschillend zijn, dat zijn de geslachtschromosomen. Afhankelijk van het geslacht beschikt men over twee X-chromosomen (vrouw) of over één X- en één Y-chromosoom (man). De chromosomen zijn weer onderverdeeld in acht groepen. Dat is gedaan aan de hand van hun grootte en hun vorm, zie Figuur 2.1. Elk chromosoom heeft een lange arm q, en bijna elk chromosoom heeft ook korte arm p. Alleen chromosomen 13, 14, 15, 21 en 22 hebben geen korte arm. Om een regio op zo'n arm aan te geven worden dan vervolgens weer cijfers gebruikt. De twaalfde regio op de lange arm van het dertiende chromosoom wordt dus genoteerd als 13q12. Van elk paar chromosomen is er één afkomstig van de moeder en één van de vader. De chromosomen bevinden zich in de kern van de cel. Ze worden alleen zichtbaar tijdens de celdeling en dan nemen ze de structuur aan die ze ook hebben in Figuur 2.1, dit wordt de metafase genoemd. Men neemt aan dat hun onzichtbare grondvorm een zeer lange, dunne draad is die door oprolling en spiraalvorming dunner en dikker wordt. Chromosomen bestaan voornamelijk uit DNA en eiwitten.
DNA is een afkorting van Deoxyribo Nucleic Acid. In het Nederlands wordt dit ook wel deoxyribonucleinzuur genoemd. DNA zorgt voor de daadwerkelijke overdracht van het erfelijk materiaal. Het bestaat uit een dubbele streng nucleotiden, die spiraalsgewijs gebonden zijn. Een nucleotide bestaat uit desoxyribose (een soort suiker), een fosfaatgroep en een van de vier nucleïnebasen. De nucleotiden verschillen alleen door die nucleïnebasen. Deze zorgen voor het erfelijk materiaal. DNA bevat namelijk de code voor de synthese van eiwitten. Een eiwit is een heel ingewikkeld molecuul, dat is opgebouwd uit (vaak vele) aminozuren. DNA bevat de volgorde waarop die animozuren aan elkaar gekoppeld dienen te worden. Drie opeenvolgende nucleotiden, een codon, zijn de code voor een bepaald aminozuur. Door een stukje DNA af te lopen wordt er een eiwit gesynthetiseerd. Een voorbeeld van zo'n eiwit is een enzym. Deze zorgen voor de stofwisseling en via reactieketens die daardoor tot stand worden gebracht komen de uiteindelijke eigenschappen van iemand tot stand. Het stukje DNA waarop precies één eigenschap ligt, wordt een gen genoemd.



Figuur 2.1: De chromosomen van een mens.

2.2 Kanker

Kanker is de naam voor een groot aantal aandoeningen die als overeenkomst hebben dat ze ongebreidelde celgroei vertonen. Door die groei verdringen de kanker cellen steeds meer gezonde cellen, waardoor hun functie achteruitgaat. Als het gezwel doorgroeit tot in een bloedvat kunnen hierdoor, soms dodelijke, bloedingen ontstaan. Tast de groei van een gezwel een zenuw aan, dan veroorzaakt dit een vaak moeilijk te behandelen pijn. De kankercellen kunnen ook uitzaaien. Daarbij laten kankercellen zich los van het gezwel en verspreiden ze zich via het bloed of de lymfe door het lichaam. Op andere plaatsen in het lichaam kunnen ze dan uitgroeien tot nieuwe gezwellen.
Kanker ontstaat doordat het DNA in een cel wordt veranderd onder invloed van bepaalde stoffen, carcinogenen genoemd, of van straling. Kankerverwekkende stoffen zijn onder andere asbest, benzeen, teer, bepaalde bestrijdingsmiddelen en zelfs sommige door andere organismen, bijvoorbeeld schimmels, geproduceerde stoffen. Bij straling moet gedacht worden aan radioactieve straling en ultraviolet licht. Er zijn twee soorten genen die met kanker te maken hebben. De één zorgt er voor dat het ontwikkelen van een tumor wordt onderdrukt, de ander bevordert juist de groei. Deze laatste wordt ook wel een oncogen genoemd. Kanker kan dus ontstaan door afname van het bij het onderdrukkende gen behorende eiwit, of door toename van het eiwit dat bij het oncogen hoort. Zo'n toe- of afname wordt veroorzaakt door een verandering in een chromosoom. Doordat er een stukje chromosoom door toedoen kankerverwekkende stoffen of straling verloren gaat, kan er een tumor onderdrukkend gen verloren gaan. Door die stoffen of straling kan het ook voorkomen dat er juist een gedeelte van een chromosoom toeneemt. Het komt dan als het ware meer voor. Als daar nu een oncogen op ligt, zijn er dan dus meerdere van. Hierdoor wordt er dus ook meer van het bijbehorende eiwit gesynthetiseerd. Verdwijnt er juist een onderdrukkend gen, dan komt juist minder van dat bijbehorende eiwit.
Veranderingen in de chromosomen werden voor het eerst ontdekt bij leukemie, bloedkanker. Door onderzoek is men er in de afgelopen 25 jaar achter gekomen welke specifieke chromosoom veranderingen er bij leukemie horen. Dat bleek bij andere soorten kanker een stuk lastiger te ontdekken. Het probleem is namelijk dat nadat er in een cel eenmaal een chromosoom verandering optreedt, er daar snel meerdere chromosoom veranderingen op volgen. Vaak hebben tumoren meer dan een dozijn chromosoom veranderingen ondergaan. Het is dan erg lastig om de uiteindelijke oorzaak nog te kunnen herleiden.

Hoofdstuk 3

Comparative Genomic Hybridization(CGH)

Verandering van de genetische structuur van een cel komt bij vele ziekten voor. In Paragraaf 2.2 is kanker al als voorbeeld genoemd. Een ander voorbeeld is het syndroom van Down. Hierbij zijn er niet twee, maar drie chromosomen nr. 21. Andere voorbeelden zijn de syndromen van Prader Willi, Angelman en Cri du Chat. Bij al deze ziektes gaat het om de vermeerdering of verlies van een chromosoom of een gedeelte daarvan. We zullen in de rest van het werkstuk voortaan de Engelse term voor vermeerdering en verlies gaan gebruiken. Bij een vermeerdering spreken we voortaan van een gain en bij een verlies spreken we voortaan van een loss. Er zijn technieken ontwikkeld om deze gains of losses te ontdekken in het DNA. Eén van deze technieken is Comparative Genomic Hybridization (CGH). CGH is ongeveer tien jaar geleden ontwikkeld en was de eerste methode waarmee in één experiment al het erfelijk materiaal van een cel kan worden gescand. In de afgelopen tien jaar heeft CGH veel aan populariteit gewonnen en zijn er belangrijke resultaten mee geboekt. In Paragraaf 3.1 zal de methode eerst besproken worden en daarna zullen de resultaten die er mee geboekt zijn in Paragraaf 3.2 worden besproken.

3.1 De methode

Er zijn verschillende methoden om CGH uit te voeren. De methode die hier behandeld wordt, is array CGH. De informatie uit deze paragraaf is afkomstig uit [6]. Verder is het één en ander uit dit artikel verduidelijkt door Kees Jong. Hij is werkzaam aan de faculteit der exacte wetenschappen van de VU, waar hij zich onder andere bezighoudt met CGH. Meer over zijn onderzoek komt terug in Paragraaf 5.1.
Bij array-CGH begint men met klonen van het DNA van gezond weefsel. Dit gekloond DNA wordt dan vervolgens verhit. Hierdoor wordt de dubbele spiraal nucleotiden verbroken. Men kiest dan vervolgens één van de beide spiralen. Dit gebeurt bij elk chromosoom. Van deze spiralen worden bepaalde regionen geselecteerd, verdeelt over alle chromosomen. Deze regionen hoeven niet op elkaar aan te sluiten en mogen elkaar zelfs overlappen. Men kiest de regionen zodanig dat ze gelijkmatig over alle chromosomen verdeeld zijn. Ook worden er regionen geselecteerd waarvan men weet dat ze belangrijk zijn. Men selecteert alleen die regionen waarvan bekend is welke genen daarop liggen. Elke regio heeft een grootte van ongeveer 100 kb. Hierbij staat kb voor kilo base, waarbij met base de nucleotidebase bedoeld wordt, en moet dus niet verward worden met kilo bytes. De regionen worden op een glazen plaat gelegd, waarbij van elke regio er 3 aanwezig zijn.
Vervolgens neemt men DNA van gezond weefsel en tumor DNA. Ook deze worden zodanig verhit dat de spiralen verbroken worden. Nu kiest men juist de andere spiraal. Deze spiralen worden in een stukjes van slechts 1 kb geknipt. Vervolgens worden ze met behulp van bepaalde moleculen verschillend gelabeld. Het DNA van het gezonde weefsel krijgt een groene fluorescerende kleur en het tumor DNA een rode fluorescerende kleur. Dit gelabelde DNA wordt dan bij het al aanwezig

DNA op de glazen plaat gevoegd. De kleine stukjes DNA zullen zich dan aan de spiraal gaan hechten van het grotere stuk DNA, op de plek waar ze oorspronkelijk ook bij hoorden. Dit wordt hybridisatie genoemd. Heeft er zich in een bepaalde regio van het DNA van de tumor nu een gain voor gedaan, dan zullen er dus meer rode dan groene stukjes aan die regio gaan hechten. Het gevolg is dat deze regio rood zal gaan kleuren. Hetgeen dus duidt op een gain in deze regio. Is er sprake van een loss in de tumor, dan zullen zich in die desbetreffende regio juist meer groene dan rode stukjes aan die regio gaan hechten. De regio zal dus groen gaan kleuren. Heeft er zich in een bepaalde regio geen verandering in het DNA voor gedaan dan zal de kleur onveranderd blijven. Er wordt software gebruikt om te analyseren waar en in welke mate er verkleuringen hebben plaatsgevonden.
Het resultaat van array CGH is een heleboel ratios. Voor elk regio wordt er namelijk een ratio berekend. Heeft het DNA in die regio geen verandering ondergaan, dan is deze ratio precies gelijk 1. Is de ratio kleiner dan een bepaalde ondergrens r^-, dan duidt dit op een vermindering. Is de ratio groter dan een bepaalde ratio r^+, dan duidt dit juist op een vermeerdering. Bij het bepalen van de r^- en r^+ houdt men rekening met de natuurlijke random verandering van DNA en met een experimentele fout. Fouten kunnen namelijk ontstaan doordat sommige stukjes DNA aan de verkeerde regio hechten. Het uiteindelijke resultaat van array CGH is dus een aantal regionen waarbij de ratio kleiner is dan de ondergrens of groter dan de bovengrens. In deze regionen zou er dus een gain of een loss plaatsgevonden kunnen hebben.
CGH heeft als voordeel dat er met één experiment al het genetische materiaal kan worden gescand op verandering in het DNA. Nog een voordeel is dat er weinig, slechts enkele nanogrammen, genetisch materiaal nodig is om CGH te kunnen uitvoeren. Het nadeel van CGH is dat de regionen waarin het DNA wordt opgedeeld nog vrij groot is. De regionen waarvan wordt aangegeven dat zich daar een gain of een loss heeft voorgedaan moeten daarom nog beter wordt onderzocht. Hiervoor gebruikt men andere technieken dan CGH, fluorescent in situ hybridization (FISH) is hier een voorbeeld van.

3.2 Resultaten en verwachtingen

Ondanks het nog korte bestaan van CGH zijn er wel al veel toepassingen in kanker onderzoek. In 1997 was er aan meer dan 1500 tumoren onderzoek verricht. Hieronder volgt een opsomming van de belangrijkste applicaties. De applicaties, als ook de andere gegevens uit deze paragraaf, zijn afkomstig uit [5].

- In de eerste plaats wordt CGH gebruikt voor het ontdekken van regionen in chromosomen waar zich een gain of een loss heeft voorgedaan. CHG is namelijk een stuk minder ingewikkeld dan de methoden die hiervoor gebruikt werden. Wanneer men gegevens van verschillende tumoren van dezelfde soort kanker combineert, is men in staat om te ontdekken welke gegevens er random zijn en welke niet. Zo wordt voor zowel borst-, eierstok-, prostaat-, nieren blaaskanker gevonden dat er zich een gain plaatsvindt in de chromosoomarmen 1q, 3q en 8q. Een loss vindt plaats in de armen 8p, 13q, 16q en 17p.

- Ook heeft CGH een belangrijke bijdrage geleverd aan kanker onderzoek door regionen aan te wijzen waar zich mogelijk kanker veroorzakende genen kunnen bevinden. Door de toepassing van CGH zijn er bij sommige soorten kanker een groot aantal regionen ontdekt waar zich veranderingen in het DNA hebben voor gedaan. Zo zijn er voor bijvoorbeeld borstkanker ongeveer 30 regionen gevonden waar zich een verandering heeft voor gedaan. Dit is veel meer dan men, op basis van onderzoek voordat CHG bestond, had verwacht.

- CGH is ook belangrijk in de studie naar de ontwikkeling van een tumor. Bij de ontwikkeling van een tumor doen zich namelijk veranderingen voor in het genetisch materiaal. Met CGH zijn deze veranderingen vooral goed te onderzoeken als men van één patiënt twee tumoren wegneemt die zich in een verschillende fase van hun ontwikkeling bevinden. Veranderingen

die zich in een eerder stadium nog niet waren voorgekomen maar in een later stadium wel, zouden kunnen duiden op regionen die belangrijk zijn bij de ontwikkeling van een tumor.

- Men heeft CGH ook succesvol weten toe te passen in het onderzoek waarbij geëxpirimenteerd wordt met modellen voor kanker ontwikkeling. Zo werden er bijvoorbeeld cellen geïnvecteerd met virussen waarvan bekend was dat ze genetische veranderingen veroorzaken. Met behulp van CGH was men in staat regionen aan te wijzen waar die veranderingen plaatsvinden.

Door deze ontdekkingen zou CGH ook gebruikt kunnen worden in de prognose en diagnose van kanker. Zo is uit onderzoek gebleken dat in 91% van de gevallen borstkanker kan worden geconstateerd door slechts te letten op drie verschillende chromosoomregionen. Het gaat hierbij dan om een gain in 1q en 8q en een loss in 13q. Hierbij komt de gain in 1q al in een beginstadium van de tumor ontwikkeling voor, terwijl de gain in 8q pas veel later optreedt. Ook kan het patroon van genetische veranderingen gebruikt worden om te voorspellen voor welk soort kanker een bepaalde chemotherapie goed werkt.
De analyses met behulp van CGH zijn eigenlijk nog in een begin stadium. Vergeleken met andere onderzoeksmethode is er nog weinig informatie beschikbaar. Met behulp van de klassieke cytogenetische analyse waren in 1998 bijvoorbeeld al 26 523 tumoren onderzocht. Ondanks dat heeft CGH al een substantieel deel bijgedragen aan het begrijpen van de ontwikkeling van kanker. Verdere studie zal meer duidelijkheid opleveren over de invloed van genetische abnormaliteiten en over de ontwikkeling van tumoren. Hierdoor zou de diagnose en behandeling van kanker patiënten wel eens kunnen gaan veranderen.

Hoofdstuk 4

Een wiskundig model voor oncogenetica

CGH levert een verzameling genetische gebeurtenissen, CNA's (copy number aberrations) genaamd. Dit zijn dus de afwijkingen die zich van normale DNA hebben voorgedaan. Deze events hebben zich in een onbekende volgorde voorgedaan. Men denkt dat deze events zich niet random voor doen, maar dat er een bepaalde samenhang bestaat. Als zich eenmaal een event heeft voorgedaan, dan verhoogt dit de kans op andere events. Door het bestuderen van hetzelfde soort type tumoren van veel verschillende patiënten zouden er patronen ontdekt kunnen worden in de genetische veranderingen. Uiteindelijk moet dit leiden tot een model voor oncogenetica, het stochastische proces van de verandering in de genetische structuur van een cel dat uiteindelijk leidt tot kanker. Zo'n model zou van grote waarde zijn bij de diagnose en behandeling van kanker.
In eerste instantie dacht men aan een pad model voor oncogenetica. Dit model is verder uitgewerkt door Vogelstein e.a., zie [8]. Vervolgens is dit model uitgebreid tot een boom model. Dit is het model dat hieronder behandeld gaat worden. Eerst zullen in Paragraaf 4.1 de soort bomen besproken worden, die we gaan gebruiken. Vervolgens zullen we voor twee bepaalde soort bomen een stelling bewijzen in Paragraaf 4.1.1. In Paragraaf 4.2 komt dan het reconstructie probleem aan de orde. Dit hoofdstuk is volledig gebaseerd op [4]. Er zullen in dit hoofdstuk heel wat symbolen worden ingevoerd. Om het werkstuk wat overzichtelijker te maken zijn de symbolen en hun betekenis ook terug te vinden in Bijlage A.

4.1 Oncogenetische Bomen

CGH levert ons een verzameling van genetische gebeurtenissen op, deze verzameling wordt gedefinieerd als V. In praktijk bevat V maximaal 82 mogelijke events. Er zijn twee soorten events. Bij het ene soort event heeft er zich een gain voorgedaan in een arm van een chromosoom en bij het andere een loss in een arm van een chromosoom. Hierbij wordt het Y-chromosoom achterwege gelaten, omdat het een erg klein en, voor het soort kanker dat wij gaan onderzoeken, onbelangrijk chromosoom is. Wij gaan namelijk in Hoofdstuk 5 borstkanker onderzoeken en daarbij is dus helemaal geen sprake van een Y-chromosoom. Verder zijn er vijf chromosomen zonder p arm en dus zijn er 41 soorten armen. Er zijn dus 2^V mogelijke deelverzamelingen van V te maken. Ook is er een kansverdeling te vinden die aan elke deelverzameling S van V een positieve kans toewijsd. Voor deze kansverdeling geldt dus dat $\sum_S p[S] = 1$.
We beginnen nu met het definiëren van verschillende soorten bomen. Allereerst de volgende gewortelde boom $T = (V, E, r)$. Hierbij is V de verzameling van alle punten van de boom, E de verzameling van alle lijnen van de boom en r is de wortel van de boom. Let op, V is hier dezelfde V als hierboven beschreven. V is dus de verzameling van alle 82 mogelijke events plus een wortel r. Voor deze boom moeten de volgende drie dingen gelden:

1. Tussen twee punten $u \in V$ en $v \in V$ bestaat ten hoogste één pad.

2. Er bestaat geen pad van $u \in V$ naar r.

3. Er zijn geen cykels.

Merk op dat de eerste eis er dus voor zorgt dat een niet samenhangende boom is toegestaan. We zijn echter wel geïnteresseerd in het gedeelte dat vanuit r bereikbaar is. Verder zorgt de tweede eis ervoor dat we hier te maken met een gerichte boom. Er zijn twee extreme gevallen van deze boom mogelijk. De ene is dat vanuit elk punt er een lijn loopt naar één ander punt, er is dan dus sprake van een pad. De ander is dat de wortel direct verbonden is met alle punten, een ster.
We kunnen dit model uitbreiden naar een gelabelde gewortelde boom, $T = (V, E, r, \alpha)$. Hierbij is $\alpha(e)$ een positief getal voor elke lijn $e \in E$. Als we eisen dat $0 < \alpha(e) \leq 1$, dan kunnen we $\alpha(e)$ als de kans dat lijn e aanwezig is. Verder nemen we aan dat er geldt dat de kans dat lijn $e_i \in E$ aanwezig is, onafhankelijk is van de kans dat $e_j \in E$ aanwezig is. We noemen $T = (V, E, r, \alpha)$ ook wel een oncogenetische boom. Deze oncogenetische bomen zijn erg handig bij het generen van de kansverdeling van 2^V. Stel nu dat S een deelverzameling is van V, waarbij S alle events bevat die bereikbaar zijn vanuit r. De kans op S kunnen we als volgt berekenen:

- Als $r \in S$, en E' is een deelverzameling van E zodanig dat S de verzameling is van alle punten die bereikbaar zijn vanuit r in de boom $T = (V, E', r)$, dan

$$p[S] = \prod_{e \in E'} \alpha(e) \cdot \prod_{(u,v) \in E, u \in S, v \notin S} (1 - \alpha(u, v)).$$

- Anders, $p[S] = 0$.

We maken dus twee aannames over oncogenetische bomen. De eerste is dat het verband tussen genetische gebeurtenissen in de vorm van een boom zijn en de tweede is dat de kans op de ene lijn onafhankelijk is van de kans op de andere lijn. Beide aannames zijn discutabel, maar zijn gedaan om het model simpel te houden en rekentijd te beperken. Het zou veel beter zijn om te veronderstellen dat het verband tussen de genetische events een gerichte acyclische graaf is. In zo'n gerichte acyclische graaf zou het zo kunnen zijn dat de kansen van de meeste lijnen laag zijn, zodat als je de lijnen neemt met een hoge kans, je een boom krijgt met de belangrijkste events. Het is verder ook te hopen dat de afhankelijkheid van de events goed kan worden benaderd door onafhankelijke kansen van de lijnen. Verder houdt het model ook geen rekening met valse positieven en valse negatieven. Hoe we dit op kunnen lossen komt in Paragraaf 4.2 aan de orde. Wat we ondanks deze beperkingen in ieder geval kunnen zeggen is dat model in ieder geval beter en uitgebreider is dan het padmodel van Vogelstein e.a., zie [8].
De oncogenetische boom die we nu hebben gedefinieerd houdt geen rekening met de factor tijd. Dit is natuurlijk wel een belangrijke factor bij oncogenetica. Daarom introduceren we hier de tijdsafhankelijke oncogenetische boom. Een tijdsafhankelijke oncogenetische boom is een gelabelde boom $T = (V, E, r, \lambda)$, samen met een positieve kansverdeling ϕ. Zo'n tijdsafhankelijke oncogenetische boom kan vervolgens in drie stappen worden gegenereerd.

1. Trek voor elke lijn $e \in E$ een getal $t(e)$ uit de exponentiële kansverdeling met gemiddelde $\frac{1}{\lambda(e)}$.

2. Trek een getal t_{tot} uit de kansverdeling ϕ.

3. Voeg vervolgens een punt v toe dan en slechts dan als er een pad P bestaat van r naar v en als er geldt dat $\sum_{e \in P} t(e) \leq t_{tot}$

Wat we nu dus doen is het volgende. Op tijdstip $t = 0$ vindt er een event plaats. Dit event wordt de wortel r van onze boom. Vervolgens geldt er voor elke lijn $e = (r, u)$, $e \in E$, dat event u een Poisson proces is met rate $\lambda(e)$. Heeft event u dan plaatsgevonden, dan geldt er voor lijn $e = (u, v), e \in E$, dat v nu een Poisson proces is met rate $\lambda(e)$. We gaan zo door totdat de som

over de $t(e)$'s voor een bepaald pad groter is dan t_{tot}. Deze t_{tot} kan namelijk gezien worden als de tijd waarop de tumor bij de patiënt verwijderd wordt. Er kunnen zich dan natuurlijk geen genetische veranderingen meer voordoen.

4.1.1 Tijdsafhankelijke of tijdsonafhankelijke oncogenetische bomen?

Het is duidelijk dat tijdsafhankelijke model veel realistischer is dan het tijdsonafhankelijk. Toch gaan we verder met het bestuderen van tijdsonafhankelijke model. Er zijn daar twee redenen voor. Ten eerste is het tijdsonafhankelijke model wiskundig beter te gebruiken dan het tijdsafhankelijke model. Neem bijvoorbeeld de kansverdeling ϕ die bij het tijdsafhankelijke model hoort. Je zou deze kunnen interpreteren als een empirische verdelingsfunctie, waarbij je voor elke tumor de tijd bekijkt tussen het begin van de genetische verandering en dat de tumor uiteindelijk werd verwijderd. Er is dan echter wel een probleem. Het is namelijk niet meer te achterhalen wanneer het begin van de genetische ontwikkeling begon. De kansverdeling ϕ wordt dan ook lastig te achterhalen. De tweede reden is dat als men eenmaal de beste tijdsonafhankelijke oncogenetische boom heeft gevonden, dat deze boom dan belangrijke aanwijzingen bevat over de beste tijdsafhankelijke oncogenetische boom. We kunnen zelfs laten dat als beide bomen een pad structuur hebben dat ze dan equivalent zijn. Hiervoor bewijzen we de volgende stelling.

Stelling 4.1 *Laat $T = (V, E, r, \lambda)$ een tijdsafhankelijke oncogenetishe boom zijn met een padstructuur en een kansverdeling ϕ. Dan is er een tijdsonafhankelijke boom T' te vinden waarvoor geldt dat deze dezelfde kansverdeling gegenereerd.*

Bewijs: Laat $T = (V, E, r, \lambda)$ een tijdsafhankelijke oncogenetishe boom zijn met de lijnen $(r, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n)$. Dit is dus een pad. We beginnen met het trekken van t_{tot} uit de kansverdeling ϕ. Laat t_i de tijd zijn waarop event v_i plaats vindt. Definieer ϕ_i als de kansverdeling van $t_{tot} - t_i$, waarbij $t_i < t_{tot}$ en $v_i \in V$. Laat verder X de verzameling van events zijn die voor t_{tot} gebeuren.
Stel e_1 is de eerste lijn van de boom, dus $e_1 = (r, v_1)$. Stel verder dat v_1 rate λ_1 heeft. De kans dat $v_1 \in X$ onder de voorwaarde dat $t_{tot} = s$ is dan gelijk aan $1 - e^{-\lambda_1 s}$, v_1 is immers een Poisson proces. Definieer nu p_1 als de kans dat $v_1 \in X$, p_1 is dan te berekenen door te integreren over de kansverdeling ϕ.

$$p_1 = P(v_1 \in X) = P(v_1 \in X | s = t_{tot}) \cdot P(s = t_{tot}) = \int_0^\infty (1 - e^{-\lambda_1 s})\phi(s)ds.$$

We hebben nu dus de kans voor de lijn (r, v_1) in termen van de tijdsafhankelijke kansverdeling ϕ. Door te integreren over s valt dus het tijdsafhankelijke deel af.
Laat nu $e_2 = (v_1, v_2)$ en λ_2 de rate behorende bij v_2. Definieer p_2 als de kans dat $v_2 \in X$ onder de voorwaarde dat $v_1 \in X$. Omdat $\phi_1(t)$ de kansverdeling is voor het interval t_{tot} onder de voorwaarde dat $v_1 \in X$, kan p_2 als volgt worden berekend

$$\begin{aligned} p_2 &= P(v_2 \in X | v_1 \in X) = P((v_2 \in X | v_1 \in X) | (s = t_{tot} - t_1 | v_1 \in X)) \cdot P(s = t_{tot} - t_1 | v_1 \in X) \\ &= \int_0^\infty (1 - e^{-\lambda_2 s})\phi_1(s)ds. \end{aligned}$$

Deze beredenatie kan natuurlijk worden uitgebreid voor het gehele pad. Definieer daarvoor λ_i als de rate behorende bij v_i, e_i als de lijn (v_{i-1}, v_i), p_i als $p(e_i)$ en ϕ_i als de kansverdeling van $t_{tot} - t_i$ onder de voorwaarde dat $v_i \in X$. Voor $i > 1$ is p_i dan als volgt te berekenen

$$p_i = P(v_i \in X | v_{i-1} \in X) = \int_0^\infty (1 - e^{-\lambda_i s})\phi_{i-1}(s)ds.$$

Op deze manier is er dus voor elke lijn een kans berekend, onafhankelijk van de tijd. We hebben dus een tijdonafhankelijke boom gekregen. □

Voordat we in Paragraaf 4.2 kunnen beginnen aan het reconstructie probleem, voeren we eerst nog enige notatie in. Een boom model, tijdsafhankelijk of tijdsonafhankelijk, heeft een kansverdeling op 2^V. De kansverdeling die bij zo'n oncogenetische boom T hoort wordt gedefinieerd als P_T. Laat P zo'n verdeling zijn en laat verder $v_i, v_j \in V$. Definieer verder S_k als de k-de deelverzameling van V, $k = 1, \ldots, 2^V$. We definiëren nu nog wat kansen. Let er op dat dit een ander soort kansen zijn dan die in het bewijs hierboven staan ook al wordt dezelfde notatie gebruikt.

- $p_i = \sum_{k=1}^{2^V} P(v_i \in S_k)$, waarbij $P(v_i \in S_k) = 0$ als v_i niet in deelverzameling S_k voorkomt.
- $p_{ij} = \sum_{k=1}^{2^V} P(\{v_i, v_j\} \in S_k)$
- $p_{i\neg j} = \sum_{k=1}^{2^V} P(v_i \in S_k, v_j \notin S_k)$
- $p_{i|j} = \frac{p_{ij}}{p_j}$
- $p_{i|\neg j} = \frac{p_{i\neg j}}{1-p_j}$

4.2 Het reconstructie probleem

We zitten nu nog wel met het volgende probleem: Gegeven een set van CGH-data, hoe construeer ik hier nu een oncogenetische boom die het beste bij de data past? Het probleem komt eigenlijk neer op hoe je bij een kansverdeling P_T de juiste boom T vindt. Voor het geval dat T een pad is, is dit vrij gemakkelijk. De kansverdeling P_T suggereert namelijk meteen hoe T moet zijn. Het event met de grootste kans moet bovenaan, dan het event met de op één na grootste kans, enz. Ook in het geval van een ster is reconstructie eenvoudig. Dit is echter in het algemene geval een stuk moeilijker. In praktijk wordt het nog ingewikkelder omdat CGH-data valse positieven en valse negatieven kan bevatten. Dat kan komen door een fout in het experiment of doordat de genetische gebeurtenissen niet relevant zijn voor de soort kanker die wordt bestudeerd. In ons model kan met valse negatieven als volgt worden omgegaan. Introduceer daarvoor bij elk punt v een nieuw punt v', wat betekent dat v is geobserveerd. De kans op de lijn (v, v') is dan gelijk aan $1 - P(valsenegatieve)$. Het is niet mogelijk om in het model op een simpele manier om te gaan met valse positieven.
In deze paragraaf zullen we zullen we een algoritme ontwikkelen om een oncogenetische boom te reconstrueren uit CGH data. In Paragraaf 4.2.1 zullen eerst beginnen met het maken van een aantal aannames die nodig zijn om het algoritme te bewijzen. Vervolgens zullen we in Paragraaf 4.2.2 het algoritme afleiden en bewijzen dat dit algoritme, bij afwezigheid van experimentele fouten en met voldoende data, de boom juist construeert. Er wordt dus bij het algoritme geen rekening gehouden met experimentele fouten en valse positieven en valse negatieven.

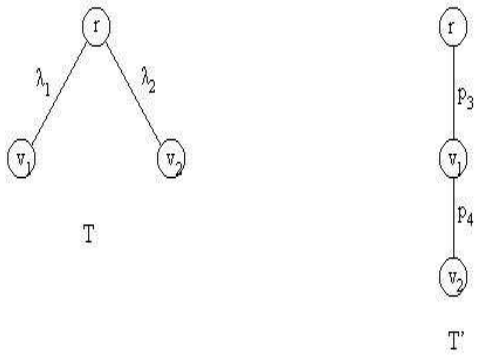
4.2.1 Enige aannamen

Om met realistische data om te kunnen gaan en om het reconstructie algoritme te kunnen bewijzen moeten we enkele aannamen maken. We zullen beginnen met een aanname te maken over de kansen van takken. Deze aanname moeten we maken om te zorgen dat events zowel onderscheidbaar als herkenbaar blijven. Stel $e = (v_i, v_j)$ is een tak van een oncogenetische boom. Staan we toe dat $\alpha(e) \to 1$ dan zijn de twee events i en j niet meer te onderscheiden. Gaat $\alpha(e)$ echter naar 0, dan kan het lastig worden om i en j, in data die niet astronomisch groot is, samen te observeren. Het moet echter wel zo zijn dat als de kans dat v_i optreedt klein is, hier ook een lage kans aan verbonden moet zijn. Als laatste moet het ook nog zo zijn dat als er een tak (v_i, v_k) bestaat, dan moeten we een redelijk aantal samples vinden waarbij v_j aanwezig was en v_k niet. Dit moet natuurlijk ook andersom gelden.
De hierboven staande problemen kunnen we als volgt oplossen. We kiezen hiervoor een constante ϵ, $\epsilon > 0$, zodanig dat $p_i > \epsilon$ voor alle i. Hierdoor is er dus een lage kans voor een event toegestaan,

maar doordat deze kans ongelijk aan 0 is voorkom je ook dat $\alpha(e) \to 0$. Verder moet voor elk paar van events (i, j) gelden dat $|p_i - p_j| > \epsilon$ of $p_i - p_{ij} > \epsilon$. Hiermee voorkom je zowel dat $\alpha(e) \to 1$, als dat er geen redelijk aantal samples voorkomen waarbij v_j aanwezig was en v_k niet, en andersom.
In een omgeving waarbij geen ruis optreedt zouden we de reconstructie als volgt kunnen doen. Hiervoor definiëren we eerst een voorouder volgorde $<_T$. Hiervoor geldt dat $v_i <_T v_j$ als $p_{i|j} = 1$, waarbij i dus de voorouder van j is. Door ons aan deze volgorde te houden kunnen we dus een boom reconstrueren. In praktijk zal deze boom echter bijna altijd een ster zijn, omdat

1. Zelfs wanneer het klopt, zullen de CNA's niet altijd in de volgorde van het model verschijnen.
2. Sommige van de CNA's die optreden zijn random.
3. Er worden door het CGH experimentele fouten gemaakt bij het rapporteren van CNA's.

Het is dus gebleken dat het in werkelijke data erg zeldzaam is dat de verschijning van een event A de verschijning van een event B strikt impliceert.
Wanneer ruis wel is toegestaan dan kan soms moeilijk worden om verschillende oncogenetische bomen te onderscheiden. Om dit duidelijk te maken kijken we naar een voorbeeld. Stel er zijn slechts twee events, v_1 en v_2. Bekijk de bomen T en T' in Figuur 4.1. We kunnen nu de parameters



Figuur 4.1: De bomen T en T'.

en de kansverdeling van de tijd zodanig kiezen dat de kansverdelingen P_T en $P_{T'}$ bijna hetzelfde zijn. Laat hiervoor $\delta > 0$ een erg klein getal zijn. Stel nu dat $P(t = t_1) = \frac{1}{2}$ en $P(t = t_2) = \frac{1}{2}$. Voor T kiezen we λ_1 en λ_2 nu zodanig dat $P(v_1 \in X | t = t_1) \geq 1 - \delta$, $P(v_1 \in X | t = t_2) = P(v_2 \in X | t = t_1) = \frac{1}{2}$ en $P(v_2 \in X | t = t_2) < \delta$. Hierdoor krijgen we voor T de volgende kansverdeling

$$\begin{aligned} p(V_1 \in X) &= p(v_1 \in X | t = t_1) \cdot p(t = t_1) + p(v_1 \in X | t = t_2) \cdot p(t = t_2) \\ &= (1 - \delta) \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4} - \frac{1}{2}\delta \\ p(V_2 \in X) &= p(v_2 \in X | t = t_1) \cdot p(t = t_1) + p(v_2 \in X | t = t_2) \cdot p(t = t_2) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \delta\frac{1}{2} = \frac{1}{4} + \frac{1}{2}\delta. \end{aligned}$$

Stel nu dat we voor T' p_3 en p_4 gelijk kiezen aan een $\frac{1}{2}$. Dan is het zo dat $P(v_1 \in X | t = t_1) = \frac{1}{2}$, $P(v_1 \in X | t = t_2) = 1$, $P(v_2 \in X | t = t_1) = 0$ en $P(v_2 \in X | t = t_2) = \frac{1}{2}$. We vinden dan voor T' de volgende kansverdeling

$$\begin{aligned} p(v_1 \in X) &= p(v_1 \in X | t = t_1) \cdot p(t = t_1) + p(v_1 \in X | t = t_2) \cdot p(t = t_2) \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} p(v_2 \in X) &= p(v_2 \in X | t = t_1) \cdot p(t = t_1) + p(v_2 \in X | t = t_2) \cdot p(t = t_2) \\ &= \frac{1}{4}. \end{aligned}$$

Wanneer ruis is toegestaan, dan zijn voor kleine waarde van δ beide bomen niet uit elkaar te houden. Dat komt doordat er geen duidelijk verschil is tussen de kans op tak (r, v_2) en de kans op tak (v_1, v_2). We voeren daarom de volgende definitie in:

Definitie: *Laat T een oncogenetische boom zijn, die de kansverdeling P_T op V genereert. T is dan schuin als geldt dat er drie punten, $v_i, v_j, v_k \in V$ te vinden zijn, zodanig dat v_k de minst algemene voorouder is van v_i en v_j en dat geldt dat $p_{i|j} \geq p_{i \vee j|k}$.*

Met de minst algemene voorouder wordt de voorouder bedoeld die v_i en v_j als eerste gemeen hebben. Dus als er vanuit zowel het punt v_i als het punt v_j het pad naar de wortel gevolgd wordt, dan is de minst algemene voorouder de eerste voorouder waar beide paden samen komen.
Een schuine boom is dus eigenlijk een boom waarvoor de kansverdeling niet overeenkomt met de topologie van de boom. Kijk bijvoorbeeld naar de boom T uit Figuur 4.1. Stel dat $p_{2|1}$ veel groter is dan $p_{r|1}$ en $p_{r|2}$ samen. Dan zou je hand van de kansverdeling verwachten dat deze boom niet de vorm had van T, maar van T'. Heeft deze boom dan toch de topologie van T dan noemen we deze boom schuin. In het tijdsonafhankelijke geval bestaan er geen schuine bomen. Stel namelijk dat v_k de minst algemene voorouder is van v_i en v_j. De kans op de lijn (v_k, v_i) is dan onafhankelijk van de kans op lijn (v_k, v_j). Hierdoor zijn de events i en j onafhankelijk van elkaar als k bekend is. Dat houdt dus in dat $p_{i|j} = p_{i|k}$. Aangezien $p_{i|k} < p_{i \vee j|k}$, zijn tijdsonafhankelijke bomen dus nooit schuin.

4.2.2 Het reconstructie algoritme

In het kort werkt het algoritme als volgt. We berekenen tussen elk paar van events een gewicht w_{ij}, waarbij in het algemeen geldt dat $w_{ij} \neq w_{ji}$. De gewichten zijn dus asymmetrisch. Vervolgens gaan we op zoek naar de gewortelde boom waarvoor het gewicht maximaal is. Deze methode wordt ook wel maximum branching genoemd. Dit is dus iets anders dan het vinden van de minimaal opspannende boom, hetgeen gedaan kan worden met het algoritme van Kruskal of het algoritme van Prim. We zullen hieronder beginnen met het definiëren van de gewicht functie. Met behulp van deze gewicht functie zullen we aantonen dat het maximum branching algoritme de boom juist construeert.
We zullen dus beginnen met het definiëren van de gewicht functie. Met behulp van deze functie kunnen van de kansverdeling van 2^V gewichten worden berekend voor de paren van events in V^2. Deze gewichten zullen we gebruiken om door middel van optimum branching de oncogenetische boom te vinden.
Intuïtief gezien moet het gewicht w_{ij} voorstellen in welke mate het gewenst is dat i de ouder van j is in de boom. In eerste instantie zou je als functie dan kunnen denken aan de likelihood ratio, $p_{ij}/(p_i p_j)$. Deze ratio is echter in ons geval ontoereikend. De gewichten moeten namelijk asymmetrisch zijn. Doordat deze asymmetrisch zijn wordt het namelijk duidelijk welke CNA er waarschijnlijk eerst optreedt. Is dus $w_{ij} > w_{ji}$ dan is het waarschijnlijker dat eerst CNA i optreedt en daarna CNA j. Als het zo is $p_i > p_j$, dan is houdt dit in dat event i vaker optreedt dan event j. Het is daarom ook beter als $w_{ij} > w_{ji}$ in plaats van andersom. Dit suggereert de volgende functie

$$w_{ij} = \frac{p_i}{p_i + p_j} \cdot \frac{p_{ij}}{p_i p_j}.$$

Willen we het reconstructie algoritme echter kunnen bewijzen dan moeten we de logaritme van de bovenstaande formule nemen. Dit leidt tot de volgende formule

$$w_{ij} = \log(p_{ij}) - \log(p_i + p_j) - \log(p_j). \tag{4.1}$$

Het nemen van de logaritme kan in sommige zeldzame gevallen leiden tot een ander optimum. Waren we in plaats daarvan op zoek naar de minimaal opspannende boom, dan had het nemen

van de logaritme de optimale boom niet beïnvloed.

Voor formule (4.1) kunnen we nu de volgende stelling bewijzen.

Stelling 4.2 *Laat T een niet schuine tijdsafhankelijk of tijdsonafhankelijk oncogenetische boom zijn. Door dan gebruik te maken van de gewichten gedefinieerd door formule (4.1) van kansverdeling P_T, leidt maximale branching over V precies tot T.*

Omdat tijdsonafhankelijke oncogenetische bomen nooit scheef zijn en vanwege Stelling 4.1 elk tijdsafhankelijk pad ook een tijdsonafhankelijk equivalent heeft, heeft Stelling 4.2 dus het volgende gevolg:

Gevolg 4.3 *Laat T een tijdsonafhankelijk oncogenetische boom of tijdsafhankelijke oncogenetische boom met een padstructuur zijn. Door dan gebruik te maken van de gewichten gedefinieerd door formule (4.1) van kansverdeling P_T, reconstrueert maximale branching over V precies T.*

We gaan Stelling 4.2 bewijzen met behulp van drie lemma's. Laat T een niet schuine oncogenetische boom zijn met wortel r en B de boom die gevonden wordt door maximum branching voor de gewichten gedefinieerd door formule (4.1). Met de drie onderstaande lemma's zullen we laten zien dat $B = T$.

Lemma 4.1 *De wortel van B is r.*

Voor het analyseren van data is r een kunstmatig toegevoegd punt. Intuïtief stelt het de cel voor zonder CNA's. Het is dus nog een gezonde cel. Er moet dus voor gezorgd worden dat het branching algoritme r als de wortel aanwijst. Dit zou dus kunstmatig kunnen, maar daardoor wordt het algoritme gecompliceerd. Door r aan te merken als een event dat bij elke tumor voorkomt, is de kans op r dus gelijk aan 1. Hierdoor retourneert het simpelere algoritme r altijd als de wortel en kunnen we Lemma 4.1 dus gebruiken.

Bewijs: Veronderstel van niet. Laat dan v_i de wortel zijn B en v_j de ouder van r in B. Het kan dus het geval zijn dat $i = j$. Bekijk nu de boom B' door in B (v_j, r) te vervangen door (r, v_i). Het verschil in gewicht van deze boom wordt dus alleen bepaald door het verschil in gewicht van deze twee lijnen. Omdat het gewicht van B maximaal is, moet dat gewicht dus groter zijn dan van B'.

$$\begin{aligned} w(B') - w(B) &= w(r, v_i) - w(v_j, r) \\ &= \log(p_{ri}) - \log(p_r + p_i) - \log(p_i) - \log(p_{jr}) + \log(p_j + p_r) + \log(p_r) \\ &= \log(1 + p_j) - \log(p_j) - log(1 + p_i) > 0. \end{aligned}$$

Dit is dus in tegenspraak met de maximaliteit van B en dus is hiermee bewezen dat Lemma 4.1 geldt. □

Laat $<_T$ de voorouder volgorde van T zijn. We hebben dit nodig in het hieronder staande lemma.

Lemma 4.2 *Laat $v_j \in V, v_j \neq r$ en laat v_i de ouder van v_j in B zijn. Dan geldt er dat $v_i <_T v_j$.*

Bewijs: Veronderstel van niet. Kies dan v_j het dichtsbij r in T met een ouder in B die in T geen voorouder is. Laat v_i die ouder in B zijn en laat verder v_k de minst algemene voorouder van v_i en v_j zijn in T. Bekijk nu de boom B' door de lijn (v_i, v_j) te vervangen door de lijn (v_k, v_j) in B. Door de keuze van v_j zullen er hierdoor geen cykels in B' ontstaan, v_j is namelijk zo dicht mogelijk bij r gekozen. Het verschil in gewicht tussen B' en B valt als volgt te berekenen

$$w(B') - w(B) = w(v_k, v_j) - w(v_i, v_j).$$

Hierbij geldt dat

$$\begin{aligned} w(v_k, v_j) &= \log(p_{kj}) - \log(p_k + p_j) - \log(p_j) \\ &= \log(p_j) - \log(p_k + p_j) - \log(p_j) \\ &= -\log(p_k + p_j). \end{aligned}$$

Er geldt dat $p_{kj} = p_j$ omdat k een voorouder van j is in T. Dus geldt er het volgende

$$\begin{aligned} p_{k|j} &= 1 \\ p_{kj} &= p_j \end{aligned}$$

We kunnen $w(v_i, v_j)$ als volgt omschrijven

$$\begin{aligned} w(v_i, v_j) &= \log(p_{ij}) - \log(p_i + p_j) - \log(p_j) \\ &= \log \frac{p_{ij}}{p_i + p_j} - \log(p_j) \\ &= \log \frac{p_{ij}}{p_j(p_i + p_j)}. \end{aligned}$$

Dus

$$\begin{aligned} w(v_k, v_j) - w(v_i, v_j) &= -\log(p_k + p_j) - \log \frac{p_{ij}}{p_j(p_i + p_j)} \\ &= \log \frac{p_j(p_i + p_j)}{p_{ij}(p_k + p_j)}. \end{aligned}$$

B is dus maximaal als

$$\begin{aligned} \log \frac{p_j(p_i + p_j)}{p_{ij}(p_k + p_j)} &< 0 \\ \frac{p_j(p_i + p_j)}{p_{ij}(p_k + p_j)} &< 1 \\ \frac{(p_i + p_j)}{(p_k + p_j)} &< p_{i|j}. \end{aligned}$$

We weten dat T een niet schuine boom is. Er geldt dan dus het volgende

$$\begin{aligned} p_{i|j} &< p_{i \vee j|k} \\ p_{i|j} &< p_{i|k} + p_{j|k} - p_{ij|k} \\ \frac{p_{ij}}{p_j} &< \frac{p_i}{p_k} + \frac{p_j}{p_k} - \frac{p_{ij}}{p_k} \\ p_{ij}(p_k + p_j) &< p_j(p_i + p_j) \\ p_{i|j} &< \frac{(p_i + p_j)}{(p_k + p_j)}. \end{aligned}$$

Hiermee hebben we dus een tegenspraak gevonden en dus geldt er dat $w(B') > w(B)$. Hiermee is Lemma 4.2 bewezen. □

Lemma 4.3 *Voor elke $v \in V$, $v \neq r$, geldt dat de ouder van v in B dezelfde ouder is van v in T.*

Bewijs: Veronderstel van niet. Laat dan v_j de ouder van $v_i \in V$ zijn in T en laat v_k de ouder van v_i in B zijn. Via lemma 4.2 geldt dan dat $v_k <_T v_j <_T v_i$. Bekijk dan de branching B die ontstaat door in B lijn (v_k, v_i) te vervangen door (v_j, v_i). Omdat $p_j < p_k$ geldt er dat

$$\begin{aligned} w(B') - w(B) &= w(v_j, v_i) - w(v_k, v_i) \\ &= \log(p_k + p_i) - \log(p_j + p_i) > 0. \end{aligned}$$

Hetgeen dus inhoud dat $w(B') > w(B)$ en dat is dus een tegenspraak, waarmee bewezen is dat Lemma 4.3 geldt. □

Met behulp van Lemma 4.1 en 4.3 is nu ook Stelling 4.2 te bewijzen. B is namelijk een opspannende boom met wortel r zodanig dat voor elke $v \in V$ de ouder van v in B ook de ouder van

v in T is (Lemma 4.3). Omdat volgens Lemma 4.1 ook geldt dat de wortel van B gelijk is aan de wortel van T, geldt er dat $B = T$. Hiermee is Stelling 4.2 dus bewezen. □

Met dit resultaat kunnen we een algoritme opstellen voor het construeren van een niet schuine boom uit samples van P_T. Dit algoritme ziet er, gegeven een set van samples van P_T, als volgt uit:

1. Bereken $\hat{p}_i$ en $\hat{p}_{ij}$, dat zijn de geschatte waarden voor p_i en p_{ij}.
2. Bereken hieruit $\hat{w}(v_i, v_j) = log(\hat{p}_{ij}) - log(\hat{p}_j) - log(\hat{p}_i + \hat{p}_j)$, de geschatte waarde voor $w(v_i, v_j)$.
3. Vindt de maximum branching B door gebruik te maken van de gewichten $\hat{w}$.

Stelling 4.4 *Wanneer T een niet schuine boom is, dan zal, mits er voldoende samples beschikbaar zijn, het bovenstaande algoritme de boom T met een hoge kans juist construeren.*

In het hieronder staande bewijs gaan we de drie bovenstaande lemma's ook voor geschatte kansen bewijzen. We zullen in dat bewijs er ook achter komen wat "voldoende" samples zijn en wat een "hoge kans" is.

Bewijs: Als eerste beginnen we met het bewijs van Lemma 4.1. We laten dus zien dat als r de wortel is van T, dat r dan ook de wortel is van B. Door de ongelijkheid uit het bewijs van Lemma 4.1 om te schrijven, vinden we dat r alleen niet de wortel van B is als er twee punten v_i en v_j zijn te vinden zodanig dat

$$\begin{aligned} \log(1+p_j) - \log(p_j) - log(1+p_i) &> 0 \\ \log(\frac{1+\hat{p}_j}{\hat{p}_j(1+\hat{p}_i)}) &\leq 0 \\ 1+\hat{p}_j &\leq \hat{p}_j(1+\hat{p}_i) \\ \hat{p}_i\hat{p}_j &\geq 1. \end{aligned}$$

Er wordt dus alleen niet aan Lemma 4.1 voldaan als zowel $\hat{p}_i = 1$ en $\hat{p}_j = 1$. Mocht er een event i bestaan waarvoor er geldt dat $\hat{p}_i = 1$, dan kunnen we dit probleem oplossen door v_i samen met r bij de wortel te groeperen. Hiermee is Lemma 4.1 bewezen voor geschatte kansen.
We zullen nu eerst laten zien dat Lemma 4.3 ook geldt voor geschatte kansen. Laat daarvoor v_k de ouder van v_i zijn in B en v_j de ouder van v_i in T, zo dat $v_k \neq v_j$. Bekijk nu de branching B' die ontstaat door uit B (v_k, v_i) te vervangen door (v_j, v_i). Dan geldt er dat

$$\hat{w}(B') - \hat{w}(B) = \log(\hat{p}_k + \hat{p}_i) - \log(\hat{p}_j + \hat{p}_i).$$

Er van uitgaande dat er op zijn minst één samples is waar v_k wel in voor komt maar v_j niet, dus $\hat{p}_k > \hat{p}_j$, dan is de bovenstaande vergelijking in tegenspraak met de maximaliteit van B. Hiermee is Lemma 4.3 bewezen voor geschatte kansen.
Als laatste bewijzen we nu Lemma 4.2 voor geschatte kansen. Veronderstel dat Lemma 4.2 niet geldt. Laat dan v_j zo dicht mogelijk bij r zijn, v_i de ouder van v_j in B, waarbij v_i geen voorouder van v_j in T is, en v_k de minst algemene voorouder van v_i en v_j in T. Definieer $\epsilon_{i|j} = p_{i \vee j|k} - p_{i|j}$. Omdat T niet schuin is, geldt er dus dat $\epsilon_{i|j} > 0$. Bekijk nu de branching B' door in B (v_i, v_j) te vervangen door (v_k, v_j). We weten dan dat

$$\begin{aligned} \hat{w}(B') - \hat{w}(B) &= \hat{w}(v_k, v_j) - \hat{w}(v_i, v_j) \\ &= \log(\hat{p}_{jk}) - \log(\hat{p}_k + \hat{p}_j) - \log(\hat{p}_{ij}) + \log(\hat{p}_i + \hat{p}_j) \\ &= \log(\hat{p}_{ik} + \hat{p}_{jk}) - \log(\hat{p}_k) - \log(\hat{p}_{ij}) + \log(\hat{p}_j) - \log(\hat{p}_k + \hat{p}_{jk}) + log(\hat{p}_k) \\ &= \log(\frac{\hat{p}_{ik} + \hat{p}_{jk}}{\hat{p}_k}) - \log(\frac{\hat{p}_{ij}}{\hat{p}_j}) - \log(1 + \frac{\hat{p}_{jk}}{\hat{p}_k}) \\ &= \log(\hat{p}_{i|k} + \hat{p}_{j|k}) - \log(\hat{p}_{i|j}(1 + \hat{p}_{j|k})) \\ &= \log(\frac{\hat{p}_{i|k} + \hat{p}_{j|k}}{\hat{p}_{i|j}(1 + \hat{p}_{j|k})}). \end{aligned}$$

Hieruit valt af te leiden dat als $\hat{p}_{i \vee j|k} > \hat{p}_{i|j}$, dat dan geldt dat $\hat{w}(B') > \hat{w}(B)$. Definieer nu $\hat{\epsilon}_{i|j} = \hat{p}_{i \vee j|k} - \hat{p}_{i|j}$. Als nu geldt dat $\hat{\epsilon}_{i|j} > 0$, dan betekent dat dat $\hat{w}(B') > \hat{w}(B)$. Hetgeen dus een tegenspraak is en daarmee is Lemma 4.2 bewezen. Het is nu alleen nog de vraag wanneer is $\hat{\epsilon}_{i|j} > 0$. Om dat te bepalen definiëren we eerst $\delta_{i|j} = \hat{p}_{i|j} - p_{i|j}$ en $\delta_{i \vee j|k} = \hat{p}_{i \vee j|k} - p_{i \vee j|k}$. We weten dan dat

$$\begin{aligned} \hat{\epsilon}_{i|j} &> 0 \\ \hat{p}_{i \vee j|k} - \hat{p}_{i|j} &> 0 \\ \hat{p}_{i \vee j|k} - \hat{p}_{i|j} + p_{i|j} - p_{i \vee j|k} &> p_{i|j} - p_{i \vee j|k} \\ \delta_{i|j} - \delta_{i \vee j|k} &< \epsilon_{i|j}. \end{aligned}$$

Deze grens gaan we zelfs wat strenger maken. We zeggen nu dat $\hat{\epsilon}_{i|j} > 0$, als geldt dat

$$\delta_{i|j} + \delta_{i \vee j|k} < \epsilon_{i|j}.$$

Voor $\delta_{i|j}$ en $\delta_{i \vee j|k}$ gaan we nu gebruik maken van de Chernoff grens. De Chernoff grens heeft drie parameters: u, deze parameter kiezen we zo, N, het aantal samples en p_{min}, de kleinste kans op een betrokken event, oftewel $p_{min} = \min_i(v_i : v_i \in X)$. De Chernoff grens is nu het volgende

$$P[\delta_{i|j} > \frac{u}{\sqrt{N p_{min}}}] < e^{-u^2/2}.$$

Als we stellen dat $u^2 = 8 \ln n$, met n het aantal punten in T (exclusief de wortel r), dan geldt er dat de rechterkant van de ongelijkheid gelijk is aan n^{-4}. We weten verder dat

$$\sum_{ij} P[\delta_{i|j} > \frac{u}{\sqrt{N p_{min}}}] > P[\max_{i,j} \delta_{i|j} > \frac{u}{\sqrt{N p_{min}}}], \tag{4.2}$$

waarbij de som en het maximum wordt genomen over de paren (v_i, v_j). Er zijn in totaal $\binom{n}{2}$ paren van (v_i, v_j), dus geldt er ook dat

$$\sum_{ij} P[\delta i|j > \frac{u}{\sqrt{N p_{min}}}] < (\frac{n^2}{2} - \frac{n}{2}) P[\delta_{i|j} > \frac{u}{\sqrt{N p_{min}}}] < (\frac{n^2}{2})(\frac{1}{n^4}). \tag{4.3}$$

Door (4.2) en (4.3) te combineren vinden we dat

$$P[\max_{i,j} \delta_{i|j} > \frac{u}{\sqrt{N p_{min}}}] < \frac{1}{2n^2}.$$

Op een zelfde wijze kunnen we afleiden dat

$$P[\max_{i,j,k} \delta_{i \vee j|k} > \frac{u}{\sqrt{N p_{min}}}] < \frac{1}{2n^2},$$

waarbij het maximum zodanig over (i, j, k) genomen wordt, dat er geldt dat v_k de minst waarschijnlijke voorouder van v_i en v_j is. Definieer nu $\epsilon = \min_{i,j} \epsilon_{i|j}$ en stel dat $\epsilon = \frac{u}{\sqrt{N p_{min}}}$, dan geldt er dat

$$N = \frac{u^2}{\epsilon^2 p_{min}} = \frac{8 \ln n}{\epsilon^2 p_{min}}.$$

We hebben nu dus gevonden wat "voldoende" en een "grote kans" is in Stelling 4.4, daarmee hebben de hieronder staande stelling bewezen.

Stelling 4.5 *Als T een boom is met n punten (de wortel r niet meegeteld), p_{min} gelijk is aan $min_i(v_i : v_i \in X)$ en ϵ gelijk is aan $min_{i,j} \epsilon_{i|j}$, dan is met $N = \frac{8 \ln n}{\epsilon^2 p_{min}}$ samples van P_T de kans dat het bovenstaande algoritme een verkeerde lijn opneemt kleiner dan $\frac{1}{n^2}$.*

Een realistisch grootte voor een boom zou 5 punten zijn, exclusief de wortel r. Als we alleen kijken naar de meest voorkomende events, dan blijkt uit de data dat 0.2 een heel plausibel waarde is voor p_{min}. Het is niet mogelijk om onze belangrijkste parameter ϵ direct te schatten. Kiezen we deze gelijk aan bijvoorbeeld 0.1, dan hebben iets meer dan 6400 samples nodig om er voor te zorgen voor elke lijn in de boom geldt dat deze correct is een kans van $24/25$. De kans dat een lijn dus fout is, is gelijk aan $1/25$. Dit komt overeen met de p-waarde van 0.05 die gebruikt wordt bij statistische toetsen. Het aantal samples is echter veel te hoog. Met de huidige CGH-techniek is een sample set die honderd maal zo klein een stuk realistischer. Het is echter te hopen dat in een boom die gebaseerd is op minder samples toch de meeste lijnen nog met een grote kans juist zijn.

Hoofdstuk 5

Analyse van VU data

In het artikel van Desper e.a., zie [4], wordt er een oncogenetische boom bepaald voor nierkanker. Dat wordt gedaan met een speciaal daarvoor geschreven programma. In het artikel staat ook dat dit programma gratis te verkrijgen is door een e-mail te sturen aan de auteurs van het artikel. Dat heb ik gedaan en kreeg een link teruggestuurd waar dit programma te vinden is, zie [2]. Met behulp van dit programma heb ik CGH-data van de VU geanalyseerd. Ik zal hieronder eerst de data beschrijven en wat ik daaruit heb geselecteerd, vervolgens zal ik de resultaten van het programma bespreken.

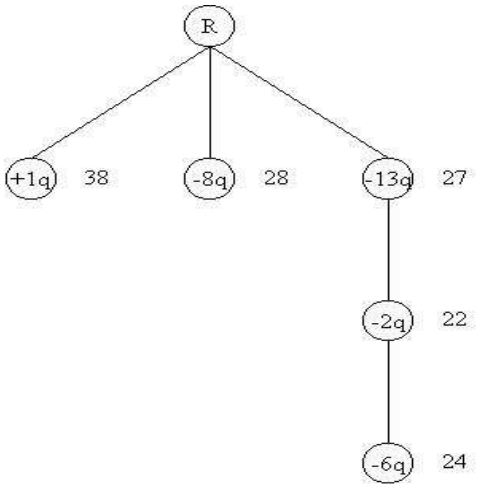
5.1 Data selectie

De CGH-data die we gebruiken in het onderzoek heb ik gekregen van Kees Jong. Hij is onder andere bezig om ruis uit de CGH-data te halen en de gains en losses te identificeren. Daarbij probeert hij te bepalen welke waarde de grenzen r^+ en r^- uit paragraaf 3.1 moeten hebben. Wanneer er in de CGH-data sprake is van een gain en wanneer van een loss is dus ook door hem voor mij bepaald. De data die Kees Jong bewerkt heeft, zijn weer afkosten van Bauke Ylstra, die werkzaam is op het medisch centrum van de VU.
De data die we hebben gekregen is van borstkanker. We hebben de beschikking over de gegevens van 54 tumoren. Elke tumor is geanalyseerd met behulp van array CGH, zoals beschreven in paragraaf 3.1. De gegevens van elke tumor zijn opgeslagen in een aparte Excel file. Zo'n Excel file heeft meer dan 2400 regels. Op elke regel staan de gegevens voor één stukje van een chromosoom, bijvoorbeeld 16p13. Over dit stukje staan een heleboel CGH gegevens, maar eigenlijk is er voor dit onderzoek maar één kolom van belang. Dat is namelijk de kolom of er sprake is van een gain of een loss. Nu is er alleen nog een probleem. Voor het computer programma willen we graag weten of er zich een gain of een loss heeft voorgedaan in een bepaalde arm en dus niet in een bepaalde regio.
Dit probleem lossen we als volgt op. Eerst gooien we de velden waarvan we weten dat er foutieve gegevens in staan weg. Zo zijn er volgens de data bijvoorbeeld chromosomen aanwezig met een nummer van 24 of hoger, terwijl maar 23 paren chromosomen bestaan. Verder wordt bij CGH het DNA van een tumor vergeleken met gezond DNA. In dit geval is er gebruik gemaakt van mannelijk referentiebloed. Aangezien we hier te maken hebben met CGH data van borstkanker, geeft dit bij het X-chromosoom verkeerde resultaten. Er treedt namelijk overal een gain op. Daarom laten we het X-chromosoom ook buiten beschouwing. Als laatste hebben we ook de lege velden weggelaten. Na de niet bruikbare velden te hebben verwijderd, selecteren we de twee benodigde kolommen. Dat is de kolom waarin staat met welke regio van het chromosoom we te maken hebben en de kolom waarin staat of er in die regio een gain dan wel een loss is.
We moeten nu gaan bepalen wanneer er sprake is van een gain of een loss in een arm van een chromosoom. Daarvoor heb ik zelf een programma in Perl geschreven. Het programma zelf is te downloaden op [3] en in Bijlage B is een handleiding voor dit programma te vinden. In dit

programma kun je instellen bij welk percentage je vindt dat er zich een gain of een loss heeft voor gedaan. Vind je bijvoorbeeld dat als meer 50% van de regionen op een arm een gain is, dat er dan sprake is van een gain op deze arm, dan stel je dit percentage dus in op 50%. Het programma geeft dan voor elke tumor als uitvoer in welke armen dat percentage wordt overschreden en of er sprake is van een gain dan wel een loss. In eerste instantie kiezen we er voor om dit percentage in te stellen op 33%. De dataset die hier uitkwomt zal in het vervolg de 33% dataset genoemd worden. De reden om het percentage in te stellen op 33% is omdat het, zoals in paragraaf 4.1 beschreven, mogelijk moet zijn om in één chromosoom, zowel een gain als een loss te kunnen observeren. Het percentage is aan de andere kant wel redelijk hoog om te voorkomen dat valse positieven invloed hebben. Er blijken nu echter enorm veel armen geselecteerd te worden, gemiddeld 17,1 per tumor. Als we de data wat nader gaan onderzoeken, blijkt dat het in deze dataset helemaal niet voorkomt dat er zowel een gain en een loss in dezelfde arm voorkomen. Er is in een arm of wel sprake van een hoog percentage gains, of wel een hoog percentage losses, of van beide een laag percentage. Een voorbeeld hiervan is te vinden in Bijlage C. Daarom kunnen we het percentage best verhogen naar 90%. Deze dataset noemen we in het vervolg de 90% dataset. Door de keuze van dit percentage hebben we geen last van valse negatieven en gaat het gemiddeld aantal armen in een tumor omlaag naar 11,3. De uitkomsten voor beide percentages zullen we in Paragraaf 5.2 analyseren met behulp van het programma dat ik gekregen heb. We zullen beide datasets gebruiken, omdat uit de verschillende datasets ook verschillende resultaten zullen blijken te komen.

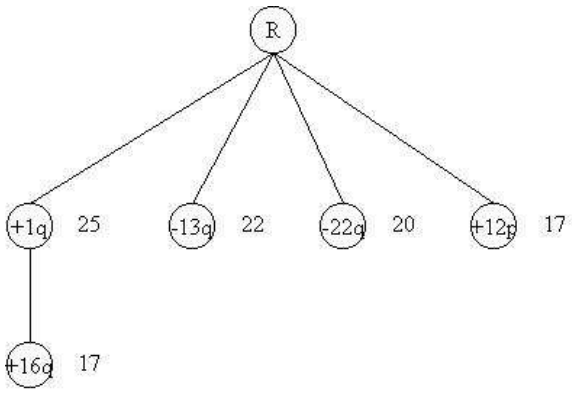
5.2 Resultaten

Het programma dat ik gedownloaded heb is een vrij groot programma. Het maken van oncogenetische boom is daar slechts een onderdeel daarvan. We zullen ook eerst van een aantal functies gebruik moeten maken voordat we het branching algoritme toe kunnen passen. Voordat dat namelijk kan moeten we eerst een voorselectie van de data maken. Er is daarvoor een speciale methode geïplementeerd, de methode van Brodeur et al. Deze kijkt er naar of er CNA's statistisch gezien abnormaal veel voorkomen. Hiervoor moeten we echter de data op een andere wijze inlezen. Het is echter niet te achterhalen hoe dat moet. In de documentatie behorende bij het programma is het niet duidelijk terug te vinden en in de broncode, die in C geschreven is, is het ook niet te vinden. Daarom kiezen we ervoor een andere methode te gebruiken. Het is namelijk ook mogelijk om aan te geven hoe groot jij je boom wilt hebben en hoeveel keer de CNA's, die in het model opgenomen kunnen worden, minimaal in de data moeten voorkomen. Op bladzijde 19 aan het eind van paragraaf 4.2.2 is aangegeven dat een aantal van 5 punten, exclusief de wortel, in de boom realistisch is. Dit aantal kiezen we dan ook voor onze boom. Verder nemen we alleen de events voor branching die 20 of meer keren in de data voorkomen. Vervolgens berekent het programma met de gewichtfunctie (4.1) de gewichten waarop de branching moet worden toegepast. Hiervoor hebben we immers in paragraaf 4.2.2 bewezen dat er dan een juiste boom gegenereerd wordt. Voor de 33% dataset ziet de boom er als volgt uit:



Figuur 5.1: De boom van de 33% dataset.

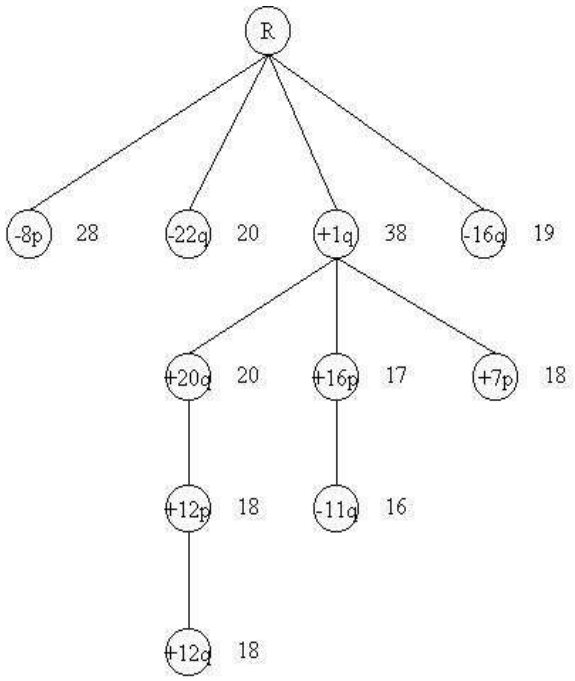
In Figuur 5.1, en ook in Figuur 5.2, 5.3 en 5.4, staat R voor de wortel, een $''+''$ voor een gain en een $''-''$ voor een loss. Het getal dat rechts van het punt in de boom staat, geeft aan hoe vaak deze bepaalde CNA in de dataset voorkomt.
In het artikel van Forozan e.a., zie [5], wordt een overzicht gegeven van welke gains en losses er bij verschillende studies gevonden zijn. De gegevens zijn afkomstig uit 1998 en in het geval van borstkanker afkomstig van in totaal 137 tumoren verdeeld over 6 studies. Van de in Figuur 5.1 voorkomende CNA's worden de gain in 1q en de losses in 8q en 13q ook in die studies gevonden.
Bij de 90% dataset komen de events logischerwijs minder vaak in de data voor. Daarom hebben we ervoor gekozen om alleen die events te gebruiken die 15 of meer keer voorkomen in de data. Voor de 90% dataset ziet de boom er dan als volgt uit:



Figuur 5.2: De boom van de 90% dataset.

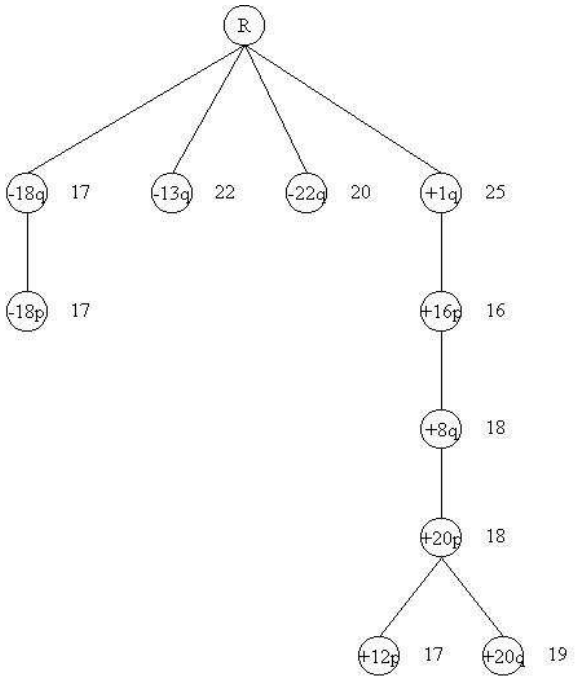
Van de CNA's in Figuur 5.2 komt alleen de gain in 12p niet in het overzicht van Forozan e.a. voor. De bomen van beide datasets verschillende zowel qua vorm als qua opgenomen CNA's. Bij de 33% dataset worden de losses in 2q en 6q opgenomen, terwijl bij de 90% dataset de gain in 12p en 16p worden opgenomen.
In paragraaf 3.2 hebben we enkele belangrijke resultaten van CGH besproken. Daar staat ook dat in 91% van de gevallen borstkanker wordt ontdekt door te letten op veranderingen in slechts drie armen. Het gaat dan om een gain in 1q en 8q en een loss in 13q. De gain in 1q en de loss in 13q

vinden we inderdaad ook Figuren 5.1 en 5.2 terug. De gain in 8q echter niet. Dat zou de volgende oorzaak kunnen hebben. Zoals ook vermeld wordt in paragraaf 3.2, treedt de gain in 1q al in het beginstadium op, terwijl de gain in 8q pas veel later optreedt. Het zou dus goed kunnen zijn dat een aantal van de tumoren al waren verwijderd voordat de gain in 8q was opgetreden. Doordat we gesteld hebben dat een gain of een loss in een arm meer dan twintig keer voor moet komen, zou het kunnen dat we de gain in 8q ten onrechte uit ons model hebben gelaten. Dat bleek voor de beide datasets echter niet het geval. In 33% dataset komt de gain in 8q 26 maal voor en in de 90% dataset 18 keer. De beperkende factor bleek het aantal punten van de boom te zijn. Het aantal punten van 5 blijkt te weinig te zijn. Daarom hebben we het aantal punten van de boom, exclusief de wortel, verhoogd naar 10. Voor de 33% dataset moet er daardoor wel een aanpassing gedaan worden. Er zijn in deze dataset namelijk maar 9 events die 20 of meer keer voorkomen in de dataset. Daarom hebben we het aantal keren dat een event voor moet komen in de 33% dataset verlaagd naar 15. Voor de 33% dataset krijgen we nu de volgende boom:



Figuur 5.3: De tweede boom van de 33% dataset

Van de events in Figuur 5.3 komen er slechts 3 niet voor in het overzicht van Forozan e.a. Dat zijn de gains in 12p, 12q en 7p. Wat opvalt is dat er een tak uit Figuur 5.1 helemaal verdwijnt. Dat is de tak met de losses in 13q, 2q en 6q. Hetgeen dus niet alleen betekent dat we door het aantal punten te verhogen voor de 33% dataset de gain in 8q niet in de boom wordt opgenomen, maar ook dat de loss in 13q uit de boom verdwenen is.
Voor de 90% dataset krijgen we het volgende resultaat:



Figuur 5.4: De tweede boom van de 90% dataset

Van de CNA in Figuur 5.4 zijn de gains in 12p en 20p en losses in 18p en 18q niet terug te vinden in het overzicht van Forozan e.a. Verder worden alle punten die in figuur 5.2 werden opgenomen, nu ook weer opgenomen. Alleen de gain in 12p verhuist naar een hele andere plek in de boom. We vinden nu ook dat de gain in 8q wordt opgenomen in de boom. Ook kunnen we uit de boom concluderen dat de gain in 1q inderdaad in een beginstadium plaatsvindt. Wat verder uit de boom in Figuur 5.4 blijkt is dat de gain in 8q ook inderdaad na de gain in 1q plaatsvindt. Ze liggen immers op dezelfde tak. Het is uit de boom echter moeilijk op te maken of de gain in 8q ook daadwerkelijk in een eindstadium plaatsvindt.

5.3 Conclusies

We hebben in de vorige paragraaf vier bomen gezien, die gebaseerd waren op onze CGH-data. Het is nu alleen nog de vraag welke is het beste. Waarschijnlijk het is zo dat een aantal van 5 punten voor een boom te weinig is in het geval van borstkanker. Zowel in de boom van Figuur 5.1 als de boom in Figuur 5.2 wordt de gain in 8q niet opgenomen. We weten echter dat deze gain belangrijk is in de ontwikkeling van kanker. Nog een reden waarom 5 punten waarschijnlijk niet realistisch is, is omdat beide bomen wel erg algemeen zijn. Ze lijken allebei erg op een ster. Ook is er uit beide bomen maar weinig informatie te halen, omdat ze vaak maar 1 punt diep zijn.
Het wordt dan dus al vrij duidelijk welke boom we als beste zullen aanduiden. In de boom in Figuur 5.3 ontbreken namelijk zowel de loss in 13q als de gain in 8q. Van beide events is bekend dat ze van belang bij de ontwikkeling van borstkanker. In de boom van figuur 5.4 worden beide events wel opgenomen, dus kunnen we concluderen dat dat model waarschijnlijk het best de genetische ontwikkeling van borstkanker gebaseerd op onze CGH-data weergeefd.
Er vallen echter nog wel wat kantekeningen te plaatsen bij de boom die we nu gekozen hebben, trouwens ook bij de andere bomen. Er vallen in alle vier de bomen namelijk twee dingen, die ik niet kan verklaren. In alle vier de bomen is het namelijk zo dat er slechts één grote tak in de boom is. De andere takken bestaan slechts in één boom uit meer dan 1 punt. Het andere dat op valt is dat in alle takken van langer dan 1 punt er, met uitzondering van één punt in Figuur

5.3, alle events van hetzelfde soort zijn. Het zijn of allemaal losses of allemaal gains. Voor beide verschijnselen heb ik geen verklaring. Ik ben natuurlijk niet echt een expert op dit gebied, maar echt logisch lijken ze me niet.
Bij het genereren van de bomen zijn er ook een aantal beperkingen. Ten eerste hebben we natuurlijk veel te weinig samples om te zorgen dat de kans dat het algoritme een verkeerde lijn opneemt klein is. Daarvoor moeten we eerst de p_{min} bepalen. Het event waarop de kleinste kans is, is natuurlijk het event dat het minst voorkomt in de dataset. Voor de 90% dataset is dat de gain in 16p. Deze komt 16 keer voor. Dus $p_{min} = 16/54 \approx 0.3$. Als we ϵ zet als in Paragraaf 4.2.2 gelijk kiezen aan 0.1, dan hebben 6140 samples nodig om er voor te zorgen dat elke tak met een kans van 99% goed wordt gegenereerd. Nu is de kans van 99% natuurlijk wel wat aan de hoge kant en kunnen we dus ook met wat minder samples af, maar onze 54 samples zijn nog lang niet genoeg. Zoals ook al eerder gezegd zijn dataset ter grootte van 6000 samples (nog) niet reëel. Het is daarom te hopen dat de bomen, en dus ook onze boom, die gegenereerd worden toch een aardig beeld geven.
Verder hebben we voor onze boom die we uiteindelijk kiezen gesteld dat er in meer dan 90 % van de regionen van een arm een gain of een loss zijn, wil er ook daadwerkelijk sprake zijn van een gain of een loss. Dit hoeft helemaal niet juist te zijn. Het zou heel goed kunnen dat er in een arm slechts een klein stukje DNA veranderd en dat dit wel degelijk van grote invloed is. Deze arm laten we dan wel mooi buiten beschouwing. In ons onderzoek wordt deze arm namelijk beschouwd als een valse positieve en wordt dus niet opgenomen. Als het slechts een paar regionen zijn, dan wordt deze ook in de 33% dataset buiten beschouwing gelaten. Om te zorgen dat het belang van die regionen wel ontdekt wordt, zouden we het experiment aan moeten passen. We zouden dan niet naar de arm van een chromosoom moeten kijken, maar naar elke regio van een chromosoom. Om dat te onderzoeken kan ook gebruik worden gemaakt van het programma dat behoord bij het artikel van Desper et al, zie [4]. Dit valt echter buiten de scope van dit werkstuk.
Als laatste hebben we nog het probleem dat we waarschijnlijk een aantal events niet in de boom opnemen die wel degelijk bij borstkanker plaatsvinden. Vooral events die pas in het eindstadium van kanker plaatsvinden hebben een grote kans om niet in de boom te worden opgenomen. Dat komt doordat er tumors al voordat zo'n eindstadium event plaatsvindt, uit de patiënt worden verwijderd. Er zijn daardoor logischerwijs minder samples met dat event. Dit event heeft dan ook een kleinere kans om in het model op te worden genomen. Het zal hierdoor waarschijnlijk moeilijk worden om echt een juiste boom voor borstkanker, of andere soorten kanker, op te stellen. Op zich is dit helemaal niet zo erg als we bedenken wat de toepassing van zo'n boom is. De boom zal worden gebruikt worden om met behulp van een CGH-experiment vast te stellen of en zo ja welk soort kanker een patiënt heeft. Een boom zal dus vooral gebruikt worden om kanker in een beginstadium vast te stellen. Het is dus vooral van belang dat in een boom dus de events die in een beginstadium plaatsvinden juist zijn. De events in het eindstadium van kanker zijn voor de toepassing dus minder van belang.
De boom in Figuur 5.4 is waarschijnlijk niet echt reëel. Toch zal het eerste gedeelte van de boom wel aardig kloppen. We vinden bij alle vier de gegenereerde bomen dat de gain in 1q in een beginstadium plaatsvindt. Dit wordt bevestigd door het artikel van Forozan e.a., zie [5]. In drie van de vier bomen vinden we ook dat er in het beginstadium losses optreden in 13q en 22q. We kunnen concluderen dat de positionering van deze drie events in het genetische proces van borstkanker het belangrijkste resultaat van ons onderzoek is.

Bijlage A

Lijst van symbolen

Hieronder volgt een lijst van symbolen en hun betekenis, die afkomstig zijn uit Hoofdstuk 4.

Symbool	Betekenis
$\alpha(e)$	De kans op lijn e.
B	De boom die ontstaat door maximum branching.
e	Een lijn van een boom
E	De verzameling van alle lijnen.
$\phi(s)$	Een positieve kansverdeling behorende bij een tijdsafhankelijke oncogenetische boom.
$\lambda(e)$	Rate van het Poisson proces behorende bij lijn e.
n	Het aantal punten in een boom, exclusief de wortel.
N	Het aantal beschikbare samples.
p_{min}	De kleinste kans op een event.
P_T	De kansverdeling van een oncogenetische boom T.
r	De wortel van een boom.
S	Een deelverzameling van V.
t_{tot}	Het tijdstip waarop het oncogetische proces eindigt.
$T = (V, E, r)$	Een gewortelde boom.
$T = (V, E, r, \alpha)$	Een gelabelde gewortelde boom of een tijdsonafhankelijke oncogenetische boom.
$T = (V, E, r, \lambda)$	Een tijdsafhankelijke oncogenetische boom.
v	Een punt van een boom.
V	De verzameling van alle mogelijke events.
X	De verzameling van alle events die voor t_{tot} plaatsvinden.
$<_T$	De voorouder volgorde van boom T, waarbij het punt links van het symbool de voorouder is van het teken rechts van het teken.

Bijlage B

Handleiding voor perl programma

Eigenlijk zijn er zelfs twee perl programma's, die ik gebruikt hebt. Het ene, script.pl genaamd, berekent per tumor in welke armen er een gain of een loss heeft voorgedaan. Het andere, berekenInput.pl genaamd, zorgt er voor dat het scipt.pl steeds met een nieuwe input wordt aangeroepen. Het programma script.pl werkt als volgt. Om te beginnen kan bovenaan in dit programma de belangrijke grens worden ingesteld wanneer er sprake is van een gain dan wel een loss. Daarna worden er een heleboel variabelen aangemaakt die later nodig zijn. Vervolgens komt het inlezen. Perl werkt zo dat het van de invoerfile steeds één regel tegelijk inleest. Er wordt dus vanuit gegaan dat de gegevens op deze regel gescheiden worden door een komma. Slechts de gegevens uit twee kolommen worden gebruikt. De gegevens over welke regio deze regel gaat en of er in die regio sprake is van een gain, een loss of dat de regio normaal gebleven is. Hierbij wordt er vanuit gegaan dat de gegevens staan zoals die stonden in de Excel files van Kees Jong. Mocht je over een andere invoer beschikken, dan is dit eenvoudig aan te passen. Wel belangrijk om dan te weten is dat de telling van de regels begint bij 0. Vervolgens werkt het programma heel simpel. Door te kijken naar de eerste twee letters van de regio wordt er bepaald tot welke arm deze regio behoort en wordt voor die regio, als daar sprake van is, het aantal gains of losses verhoogd. Ook wordt het totaal aantal regio's dat van een arm is onderzocht bijgehouden. Is de hele file ingelezen dan worden de armen uitgeprint waarbij het percentage gains en losses hoger was dan de ingestelde waarde. Eerst worden de gains, voorafgegaan door een ?, uitgeprint, vervolgens worden de losses, voorafgegaan door een !, uitgeprint. Op deze manier is namelijk de input voor het programma van Desper e.a. Het programma berekenInput.pl zorgt er dus voor dat dit voor alle files gebeurd. Daarvoor moeten wel alle namen van deze files in de broncode van het programma worden ingevoerd. De aanroep van dit programma werkt onder Unix op VU als volgt

```
perl berekenInput.pl
```

De input is nu nog niet helemaal goed voor het programma van Desper et al. Daarvoor moet er voor elke regel nog $p1, p2, \ldots$ komen te staan. Dit kan het beste gedaan worden door zon kolom even in Excel er voor te plakken.

Bijlage C

Aantal gains en losses in een tumor

In de hier onderstaande tabel staat een overzicht van het resultaat van CGH van één onderzochte tumor. Bij elke arm staat aangegeven in hoeveel regionen van die arm een gain of een loss heeft plaatsgevonden en het totaal aantal onderzochte regionen.

Arm	Gains	Losses	Totaal
1p	13	38	76
1q	60	1	62
2p	0	0	24
2q	0	26	43
3p	0	38	44
3q	1	14	51
4p	0	45	45
4q	18	87	131
5p	26	0	26
5q	16	8	88
6p	48	0	48
6q	5	0	39
7p	0	56	69
7q	36	78	115
8p	5	38	66
8q	90	0	90
9p	0	25	41
9q	0	25	71
10p	42	0	42
10q	93	0	93
11p	42	0	42
12p	22	0	22
12q	34	0	67
13q	0	0	56
14q	0	0	71
15q	0	60	69
16p	12	0	16
16q	42	0	42
17p	0	0	16
17q	36	0	42

Arm	Gains	Losses	Totaal
18p	0	0	19
18q	0	0	35
19p	0	0	13
19q	0	0	24
20p	32	0	32
20q	43	0	43
21q	0	0	34
22q	17	0	17
Xp	22	0	22
Xq	32	0	33

Bibliografie

[1] 1998, Encarta 98 Encyclopedie - Winkler Prins Editie, Ahaed Software, Duitsland.

[2] http://fastlink.nih.gov/pub/staff/schaffer/oncogenesis

[3] http://www.cs.vu.nl/~jvhooff/werkstuk

[4] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou & A. Schäffer, *Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data*, Journal of Computional Biology 6 (1999), blz. 37-52

[5] F. Forozan, R. Karhu, J. Kononen, A. Kallioniemi & O. Kallioniemi *Genome screening by comparative genomic hybridisation*, Trends In Genetics 13 (1997), blz. 405-409.

[6] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, Y. Zha, S. Dairkee, B. Ljung, W. Gray & D. Albertson, *High resolution analysis of DNA copy number variation using comparative genomic hybridisation to microarrays*, Nature Genetics 20 (1998), blz. 207-211.

[7] E. van de Schoot, *VWO Biologie Samengevat* Onderwijspers BV (1997), Leiden, blz 202-221.

[8] B. Vogelstein, E. Fearon, S. Hamilton, S Kern, A. Preisinger, M. Leppert, Y. Nakamura, R White, A. Smits & J. Bos, *Genetic Alterations During Colorectal Tumor Development*, New England Journal of Medecine 319 (1988), 525-532.