Toepasbaarheid van Bayesiaanse Netwerken voor Cellulaire Signaal Transductie

door

Mart van Rijthoven, 3588262
11 Juli 2016

Eerste beoordelaar: Silja Renooij
Tweede beoordelaar: Gerard Vreeswijk
Opleiding: Bachelor Kunstmatige Intelligentie, Universiteit Utrecht
Ects: 7.5

Samenvatting

In deze scriptie bespreek ik de toepasbaarheid van Bayesiaanse netwerken voor cellulaire signaal transductie. Ik leg uit wat Bayesiaanse netwerken zijn en wat cellulaire signaal transductie inhoudt. Ik laat zien hoe Bayesiaanse netwerken kunnen worden toegepast bij onderzoek naar cellulaire signaal transductie en wat de toegevoegde waarde is van deze toepassing.

Trefwoorden

Bayesiaans, netwerken, grafen, cellen, signaal transductie, leren, kunstmatige intelligentie.

Inhoudsopgave

Hoofdstuk 1

Inleiding

Cellulaire signaal transductie zijn moleculaire interacties die processen in het lichaam reguleren. Om deze interacties te voorspellen kan er gebruik gemaakt van Bayesiaanse netwerken. Dit is een tool die gebruikt wordt in de kunstmatige intelligentie. Bayesiaanse netwerken kunnen probabilistische relaties representeren. De netwerken kunnen erg belangrijk zijn in de biomedische wetenschappen voor het ontdekken en voorspellen van moleculaire interacties [5]. Door cellulaire signaal transductie beter te begrijpen kunnen er medicijnen voor verschillende ziektes worden ontwikkeld.

Ik schrijf over dit onderwerp omdat ik de samenwerking van Bayesiaanse netwerken en biomedische wetenschappen interessant vind. Ik denk de toepassing van Bayesiaanse netwerken nieuwe inzichten kan geven over biomedische wetenschappen die nog niet bekend zijn.

De onderzoeksvraag die ik ga beantwoorden in deze scriptie is:

- Hoe kunnen Bayesiaanse netwerken toegepast worden bij onderzoek naar cellulaire signaal transductie en wat is de toegevoegde waarde?

Daarbij beantwoord ik de vragen:

- Wat voor soort Bayesiaanse netwerken kunnen worden gebruikt voor signaal transductie?
- Wat voor soort signaal transductie data is er beschikbaar om te gebruiken?
- Kunnen Bayesiaanse netwerken eiwit-interacties voorspellen? En hoe bruikbaar zijn deze voorspellingen?

- Kunnen Bayesiaanse netwerken cellulaire cascades modelleren? En hoe goed passen deze modellen op de consensus modellen?

Deze vragen ga ik beantwoorden door naar verschillende toepassing van Bayesiaanse netwerken te kijken die gebruikt zijn in het domein van signaal transductie. Om mijn onderzoeksvraag te beantwoorden onderzoek ik wat Bayesiaanse netwerken zijn en wat er mee gedaan kan worden. Dit bespreek ik in het volgende hoofdstuk. Daarnaast onderzoek ik wat signaal transductie inhoudt. Dit leg ik uit in hoofdstuk 3. In de jaren tussen 2000 en 2010 zijn er artikelen verschenen waarin Bayesiaanse netwerken zijn toegepast bij onderzoek naar cellulaire signaal transductie. In hoofdstuk 4 bespreek ik deze artikelen. In het laatste hoofdstuk zal ik de literatuur kritisch beoordelen en mijn bevindingen presenteren. Daarnaast zal ik concluderen onder welke omstandigheden Bayesiaanse netwerken toegepast kunnen worden en een toegevoegde waarde zullen hebben bij onderzoek naar cellulaire signaal transductie.

Hoofdstuk 2

Bayesiaanse Netwerken

In dit hoofdstuk bespreek ik wat Bayesiaanse netwerken zijn. Omdat Bayesiaanse netwerken onder andere bestaan uit stochastische variabele, bespreek ik in het kort hoe kansen zijn gedefinieerd en welke notatie ik gebruik. Ik bespreek de relevante technieken die bij deze netwerken worden gebruikt in het domein van cellulaire signaal transductie. Dit zijn inferentie, model selectie en model ontdekken. Daarnaast zal ik uitleg geven over dynamische Bayesiaanse netwerken. Dit is een speciaal geval van een Bayesiaans netwerk dat met tijd en feedback systemen kan omgaan. In dit hoofdstuk heb ik gebruik gemaakt van het boek: Artificial Intelligence: A Modern Approach [7] tenzij anders aangegeven.

2.1 Waarschijnlijkheid

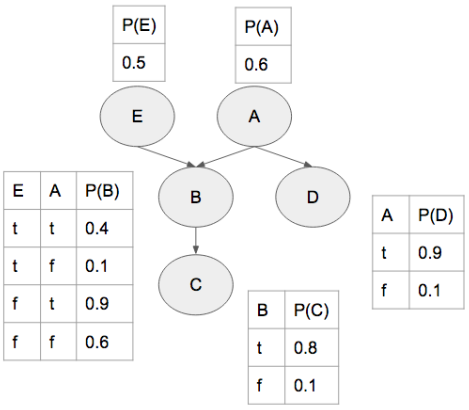
Onzekerheid ontstaat door gedeeltelijke observaties, non-determinisme en gebrek aan kennis. Om hiermee om te gaan kan er gebruik gemaakt worden van waarschijnlijkheid. In de waarschijnlijkheidsleer wordt er gesproken over mogelijke werelden. Een wereld is een verzameling van stochastische variabelen aangeduid met een hoofdletter. Elke stochastische variabele heeft een domein van mogelijke waardes. Elke waarde in dit domein heeft een kans aangeduid met P(a), waar a staat voor een waarde uit een domein. Alle kansen van een stochastische variabele A tezamen, noemen we de waarschijnlijkheidsdistributie van A, genoteerd als P(A). De volledige gezamenlijke waarschijnlijkheidsdistributie wordt gerepresenteerd door alle combinaties van alle stochastische variabelen.

Een kans op een waarde in het domein van een stochastische variabele kan onvoorwaardelijk (a-priori) zijn. Dit is een kans zonder rekening te houden met andere informatie. Een voorwaar-

delijke kans (a-posteriori) is de kans op een waarde uit het domein gegeven een andere waarde uit een domein genoteerd als: P($a|b$). Voorwaardelijk kansen kunnen gedefinieerd worden in termen van onvoorwaardelijke kansen: P($a|b$) = $P(a \wedge b)/P(b)$. Twee variabelen zijn onafhankelijk wanneer geldt dat P($a \wedge b$) = P(a)P(b), voor alle waarde van a en b. En wanneer dus geldt dat P($a|b$)=P(a).

2.2 Bayesiaanse Netwerken

Een Bayesiaans netwerk (BN) is een verzameling van stochastische variabelen met hun conditionele afhankelijkheden gerepresenteerd met een gerichte a−cyclische graaf. De knopen in de graaf stellen stochastische variabelen voor. De pijlen geven de directe afhankelijkheden weer tussen de stochastische variabelen. Als er een pijl is van X_i naar X_j, dan heet X_i de ouder van X_j. Elke knoop X_i heeft verzameling voorwaardelijke kansverdelingen $P(Xi|Parents(Xi))$ voor elke combinatie van waarden voor de ouders, die het effect van de ouders op de knoop bepaalt. Wanneer de voorwaardelijke kansverdeling voor elke variabele gegeven zijn ouders is gespecificeerd, kan hiermee de volledig gezamenlijke waarschijnlijkheidsdistributie gespecificeerd worden.



Figuur 2.1: Een simpel Bayesiaans netwerk. Bij de kansverdeling staat t voor true en correspondeert met activatie. De f staat voor false en correspondeert met inactivatie. Complementaire kansen zijn niet weergegeven. Gebaseerd op *Bayesian Network Analysis of Signaling Networks: A Primer [6]*

In figuur 2.1 is een simpel Bayesiaans netwerk weergegeven met 5 knopen: A,B,C,D,E. De fictieve kansverdeling is weergeven in een kansverdeling tabel ter illustratie. In dit voorbeeld stellen de knopen activatoren/in−activatoren voor. A heeft een pijl naar B en zal ervoor zorgen dat B geactiveerd wordt. De verwachting is dan dat wanneer A geactiveerd is B ook actief zal zijn. Dus kennis over A geeft informatie over B. Vanuit B is er een pijl naar C. B zal C activeren. Wanneer A actief is, zal de verwachting zijn dat C ook actief is, door de activatie van B. Maar als we weten dat B actief is, dan zal de kennis over de toestand van A ons verder geen informatie geven. A en C zijn dus voorwaardelijk onafhankelijk gegeven B. A heeft ook een pijl naar D zodat A ook D activeert.

Wanneer B actief is kan dit veroorzaakt zijn door A. Omdat A ook D activeert kan er geredeneerd worden dat wanneer B actief is D ook actief kan zijn. B en D zijn dus ook voorwaardelijk onafhankelijk gegeven A. Dan is er ook nog een pijl van knoop E naar B. E kan B inactiveren. B wordt gereguleerd door A en E. Om nu iets te kunnen zeggen over de activatie van B kan er gekeken worden naar de voorwaardelijke kansverdeling van B geven beide ouders A en E [6].

2.2.1 Inferentie

Inferentie is de berekening van a−priori of a−posteriori kansen voor een gevraagde variabele gegeven een verzameling van geobserveerde variabelen. Gegeven de gevraagde kansen: P($X|e$) kan deze gevalueerd worden door:

$$P(X|e) = P(X,e)/P(e) = \alpha P(X,e) = \alpha \sum_{y} P(X,e,y),$$

waar X de gegeven discrete variabele is, e de verzameling van geobserveerde discrete variabelen en y de niet geobserveerde discrete variabelen. α is een normalisatie constante die ervoor zorgt dat de verdeling optelt tot 1.

2.2.2 Continue variabelen

Het voorbeeld in figuur 2.1 gaat uit van discrete waarden, namelijk geactiveerd of geïnactiveerd. Meestal zijn de variabelen in de werkelijkheid continu. De variabelen hebben dan een oneindig aantal mogelijke waarden. Het is onmogelijk om aan al deze waarden een conditionele kans toe te wijzen. Er zijn verschillende mogelijkheden om toch met continue variabelen om te gaan. De variabelen kunnen bijvoorbeeld gediscretiseerd worden. Hierbij worden waarden opgedeeld in een verzameling van intervallen. Een andere mogelijkheid, en vaak gebruikt, is het gebruik van de normale verdeling:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

2.2.3 Leren

Parameters leren

Om de kansverdeling van X te leren voor een netwerk met een gegeven structuur, kan er uit training data D de parameters θ geleerd worden zodanig dat het de waarschijnlijkheid (L) maximaliseert dat de data van X komt [5].

$$L(\theta) = p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

Model ontdekken

Om een model te ontdekken kunnen modellen gegenereerd worden. Een methode is om willekeurig een pijl in het netwerk toe te voegen, te verwijderen of de pijl tussen variabelen om te draaien. Wanneer de structuur veranderd is kan het model opnieuw gescoord worden. Als het nieuwe model beter scoort wordt de verandering behouden. Om lokale minimum te vermijden kan er met een kleine kans veranderingen opgenomen worden die het netwerk verslechteren [8].

Model selectie

Verschillende modellen kunnen gescoord worden tegen data om de waarschijnlijkheid te bepalen tussen de modellen. De Bayesiaanse score van een model S is de voorwaardelijke kans dat model S de juiste is gegeven de data D. Dit kan dan gebruikt worden om de verschillende modellen te scoren. Model selectie kan worden toegepast met de volgende Bayesiaanse score:

$$Score(S) = log\, P(S|D)$$

2.2.4 Structuren

Een naïeve Bayesiaans netwerk is een simpel netwerk met een classificatie knoop en andere knopen die alle conditioneel onafhankelijk van elkaar zijn. Door de parameters voor het netwerk te leren kan er geclassificeerd worden. Het netwerk wordt als volgt gedefinieerd: $P(C|X_i..X_n)$, met C de classificatie node en X_i de variabelen die kenmerkend zijn voor de classificatie. Volledige Bayesiaanse netwerken zijn netwerken waar elke knoop alle mogelijke ouders als ouder heeft. Bayesiaanse netwerk structuren kunnen ook gedefinieerd zijn met behulp van expert kennis. De pijlen worden dan bepaald door experts in het domein waar het Bayesiaanse netwerk in wordt toegepast.

2.3 Dynamische Bayesiaanse Netwerken (DBNs)

Een dynamisch Bayesiaanse Netwerk (DBN) is een Bayesiaans netwerk met tijds intervallen. De simpelste vorm van een DBN is een Hidden Markov model (HMM). In een HMM is de toestand beschreven met 1 discrete variabele. DBNs kunnen meerdere variabelen hebben. Elk interval kan een aantal toestandsvariabelen hebben aangegeven met X_t en de gegeven geobserveerde variabelen E_t. Om een DBN te construeren zijn er nog twee andere modellen nodig: het transitie model en het sensor model. Het transitie model specificeert de kansverdeling over de huidige

toestand variabelen gegeven de vorige waardes:

$$P(X_t|X_{0:t-1})$$

De verzameling $\{X_{0:t-1}\}$ is oneindig in de tijd t. Dit is een probleem omdat er geen oneindige kansverdelingen kunnen worden gespecificeerd. Wanneer er een Markov assumptie wordt gemaakt worden niet alle toestanden in beschouwing genomen maar kan er worden gekeken naar een eindige reeks van vorige toestanden. Er kan bijvoorbeeld gekeken worden naar alleen de vorige toestand. Het transitie model wordt dan als volgt gedefinieerd:

$$P(X_t|X_{0:t-1}) = P(X_t|X_{t-1})$$

Omdat t ook een oneindig aantal waarden kan aannemen kan er aangenomen worden dat $X_{0:t-1}$ voor elke t dezelfde voorwaardelijke kans heeft.

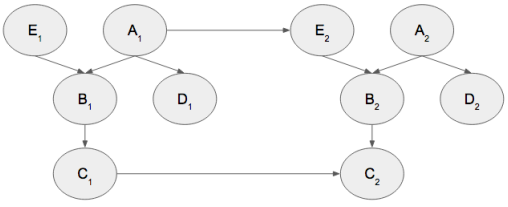
Naast het transitie model is er ook een een sensor model nodig. De gegeven geobserveerde variabelen kunnen afhankelijk zijn van vorige variabelen en ook van de huidige verzameling van toestandsvariabelen. Als we een sensor Markov assumptie maken is het sensor model als volgt gedefinieerd:

$$P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t)$$

Naast het specificeren van het transitie en sensor model moet de a−priori kansverdeling op tijd t=0 gegeven zijn: $P(X_0)$. Met deze ingrediënten kan als volgt de volledige gezamenlijk distributie gespecificeerd worden over alle variabelen voor elke t:

$$P(X_{0:t}, E_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i|X_{i-1})P(E_i|X_i)$$

In Figuur 2.2 is een simpel dynamische Bayesiaans netwerk gezien. Hierin is E tijdsafhankelijk van A. C tijdsafhankelijk van zichzelf. De regulatie van C op zich zelf kan gezien worden als feedback.



Figuur 2.2: Een simpel dynamische Bayesiaans netwerk. De subscripts staan voor t de tijd. Gebaseerd op *Bayesian Network Analysis of Signaling Networks: A Primer [6]*

Hoofdstuk 3

Cellulaire Signaal Transductie

Cellulaire signaal transductie is het doorgeven van signalen op cel niveau met als doel regulatie van cellulaire processen. In dit hoofdstuk bespreek ik in het kort wat dit precies inhoudt. Eerst beschrijf ik wat cellen zijn. Daarna bespreek ik essentiële componenten in de signaal transductie. Hierop volgt een korte sectie over moleculaire technieken die gebruikt worden om deze componenten te onderzoeken. Daarna beschrijf ik wat een signaal route inhoudt en laat ik een consensus route zien uit de literatuur. Als laatst bespreek ik signaal transductie tussen twee naastgelegen cellen en de extracellulaire matrix. Voor dit hoofdstuk heb ik gebruik gemaakt van het boek: *Moleculair biology of the cel*[1] en het volgt voornamelijk hoofdstuk 15 uit dit boek.

3.1 Cellen

Cellen zijn dynamische eenheden die kunnen overleven, groeien, delen, differentiëren en dood gaan. Een cel is van zijn omgeving afgesloten door een celmembraan. Het gebied binnen de cel wordt intracellulair genoemd, het gebied dat buiten de cel ligt wordt de extracellulaire ruimte genoemd. Een cel bevat organellen, dit zijn compartimenten met een eigen celmembraan. Twee belangrijke organellen zijn de kern en het endoplasmatisch reticulum (ER). De kern bevat al het genetische materiaal en hier vindt de productie van mRNA plaats. Aan het ER vindt er een vertaling plaats van mRNA naar een keten aminozuren, oftewel een eiwit.

3.2 Eiwitten

Eiwitten reguleren veel processen in de cel. Eiwitten zijn modulair opgebouwd uit verschillende domeinen. Verschillenden domeinen kunnen binden aan verschillende aminozuren.

3.2.1 Eiwit-Eiwit Interacties

Eiwit-eiwit interacties zijn belangrijk bij signaal transductie. De interacties zijn de bouwstenen van de cascades. Door willekeurige bewegingen van eiwitten in een cel zullen eiwitten elkaar tegenkomen en tegen elkaar aan botsen. Als beide eiwitten de juiste conformatie hebben dan kunnen ze aan elkaar binden. Hierdoor kan er een conformatie verandering optreden. Door de verandering kan er bijvoorbeeld een actieve plek ontstaan waar een ander molecuul weer aan kan binden. Eiwitten die aan elkaar verbonden zijn worden eiwit complexen genoemd.

Activatie en Inhibitie

Een enzym is een eiwit dat als katalysator fungeert. Enzymen die een belangrijke rol spelen in de signaal transductie zijn kinases en fosfatases. Deze enzymen kunnen eiwitten activeren en inactiveren door een fosfaat op het eiwit te zetten of juist er af te halen. Een guanine nucleotide−bindings eiwit (g−eiwit) heeft een soort gelijk systeem en kan dus ook eiwitten activeren en inactiveren.

Receptoren

Receptoren zijn eiwitten waar signaalmoleculen aan kunnen binden. Er bestaan verschillende soorten receptoren. Een receptor kan zich transmembraan bevinden. Dit houdt in dat de receptor een extracellulair en intracellulair gedeelte bevat. Daarnaast zit er een gedeelte vast in het membraan. Dit is belangrijk omdat het extracellulaire deel voor zorgt voor binding van een signaal molecuul en het intracellulair gedeelte signalerend is binnen de cel. Een ionkanaal, wat ook een transmembraan receptor is, kan een signaal waarnemen en kan het kanaal laten open gaan. Hierdoor kunnen er ionen door het kanaal stromen. Een g-eiwit gekoppelde receptor (GPCR) is een receptor die gebonden is aan een g-eiwit. Wanneer een signaal extracellulair bindt aan de receptor zal het g-eiwit actief worden en kan er intracellulair een signaal cascade in werking worden gesteld.

Second Messengers

Second messengers zijn moleculen en brengen het signaal dat is ontstaan bij de receptor door in de cel en zijn essentieel in signaal transductie cascades. Enkele belangrijke second messengers zijn: cyclische AMP (cAMP), Ca^{2+}, Inositol-1,4,5 trifosfaat (IP3), Diaglycrol (DAG), Phosphatidylinositol difosfaat (PIP[4,5]2).

3.2.2 Technieken

Er bestaan verschillende technieken om aan te tonen of een eiwit wel/niet of veel/weinig aanwezig is in een cel. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS−PAGE) is een techniek die alle eiwitten in een cel zichtbaar kan maken aan de hand van een moleculaire ladder. Met Western blotting kan door middel van een antilichaam een specifiek eiwit zichtbaar worden gemaakt. Immunoprecipitatie maakt ook gebruik van antilichamen maar kan worden toegepast wanneer het eiwit in mindere mate aanwezig is. Met flow cytometrie kan expressie van meerdere eiwitten, tegelijk gemeten worden in duizenden individuele cellen.[6]

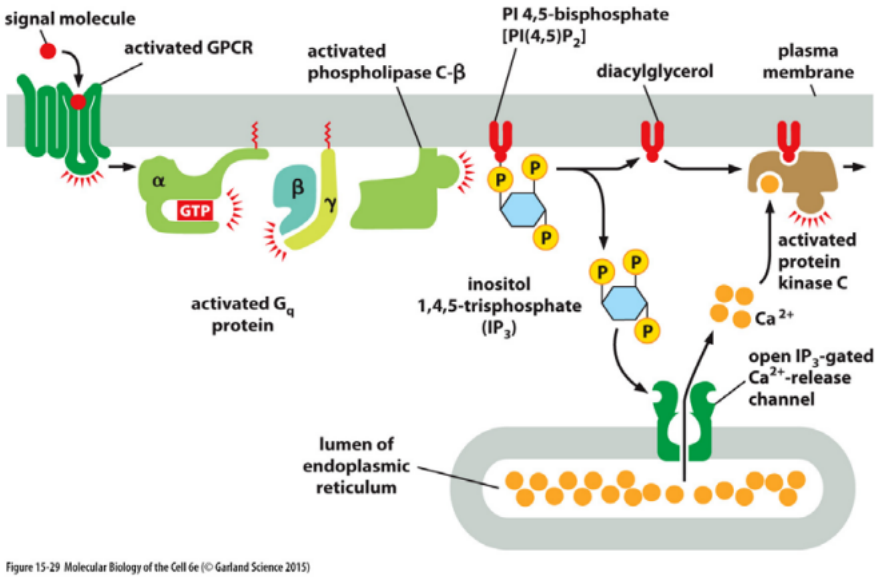
Om aan te tonen dat een eiwit actief is kan er met fosfo-antilichamen gekeken worden of het gefosforyleerde eiwit, een eiwit met een fosfaat er aan, aanwezig is. Daarnaast is het mogelijk activiteit af te leiden door naar het effect te kijken wanneer er mutaties aan het eiwit aangebracht worden. Ook kunnen er activatoren en inhibitoren in de cel worden gebracht en kan er gekeken worden naar het verwachte effect. Om de locatie en translocatie van eiwitten in beeld te brengen kunnen eiwitten fluorescent gelabeld worden. Eiwit-eiwit interactie kan aangetoond worden door co-immunoprecipitatie waarbij er antilichamen tegen een eiwit uit het complex wordt gebruikt.

3.3 Signaal Transductie

Signalen komen van buiten de cel en worden doorgegeven binnen de cel. De cel zal hierdoor een bepaalde respons geven. Een signaal transductie netwerk reguleert de tijd en plaats van eiwitten. Hierbij zijn positieve en negatieve regelsystemen betrokken.

3.3.1 Signaal Routes

In Figuur 3.1 is een voorbeeld van een signaal route in cartoon-style weergegeven. Dit betreft een cascade die voor de activatie van protein kinase C (PKC) zorgt. Door binding van een signaal molecuul aan een GPCR zal de receptor actief worden. De receptor zal het gebonden g-eiwit activeren waarna de beta−subeenheid van dit g-eiwit fosfolipase c (PLC) activeert. PLC zal vervolgens PIP[4,5]2 knippen in DAG en IP3. IP3 bindt aan een ionenkanaal en zorgt dat deze opengaat. Dit ionenkanaal zit in het membraan van het ER. Het lumen van het ER bevat meer calcium dan de cel zelf. Door de opening en het concentratie verschil zal calcium de cel in diffunderen. DAG zal binden aan PKC, het vrijgekomen calcium zal ook aan PKC binden waardoor PKC actief wordt. De PKC route is een simpel voorbeeld van de vele routes die in de



Figuur 3.1: PKC route, Moleculair biology of the Cell [1]

cel bestaan. In hoofdstuk 4 worden er ook andere routes besproken

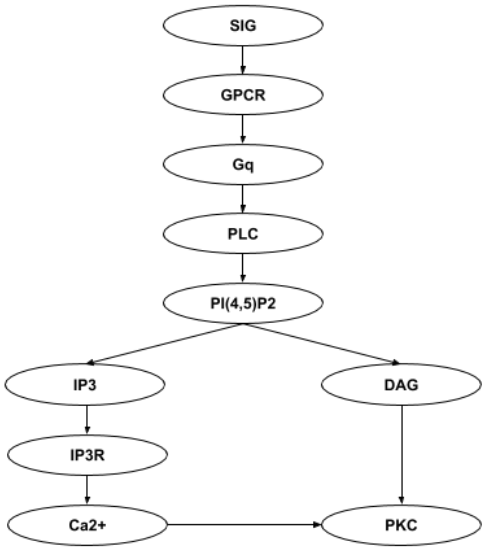
3.3.2 Celadhesie

Naast dat signalen cascades in individuele cellen in gang kunnen zetten is het ook van belang dat bepaalde cellen verbonden zijn aan elkaar of aan een extracellulaire matrix. Dit doen cellen door middel van verbindingen (celadhesie). Bij huidcellen is het bijvoorbeeld van belang dat de cellen aan elkaar blijven. Belangrijke moleculen hiervoor zijn cadherines en integrines. Cadherines zorgen voor adhesie tussen naastgelegen cellen. Integrines zorgen voor adhesie tussen de cel en de extracellulaire matrix. De cadherines en integrines bevinden zich transmembraan. Hierdoor kunnen cellen bijvoorbeeld bij elkaar blijven of juist gaan differentiëren.

Hoofdstuk 4

Bayesiaanse Netwerken toegepast voor Signaal Transductie

In dit hoofdstuk bespreek ik literatuur waar Bayesiaanse netwerken zijn toegepast in het domein van cellulaire signaal transductie. De signaal cascades die we hebben gezien in het vorige hoofdstuk hebben verschillende componenten die invloed op elkaar kunnen hebben. De afhankelijkheden in deze cascades kunnen worden gerepresenteerd in Bayesiaanse modellen. De knopen zijn componenten en de pijlen weergeven de afhankelijkheden tussen de componenten. In Figuur 4.1 is een graaf weergegeven dat als model staat voor de PKC route die is besproken in het vorige hoofdstuk. De knopen zijn de componenten uit de cascades. Het signaal (SIG), een receptor-bindend eiwit, is het eerste component uit de PKC route en dus ook de eerste knoop in het Bayesiaans netwerk. Verder volgt het netwerk de structuur van het model gegeven in figuur 3.1. We hebben gezien dat PKC zowel calcium (Ca^{2+}) als diacylglycerol (DAG) nodig heeft om in een actieve toestand te komen. En de aanwezigheid van Ca^{2+} en *DAG* worden beide gereguleerd door PIP[4,5]2. Dus de verwachting is dat wanneer PIP[4,5]2



Figuur 4.1: Een persoonlijke interpretatie van een Bayesiaans netwerk voor de PKC route

zich splitst in IP3 en DAG, PKC actief zal worden. Maar gegeven DAG en IP3 is PKC conditioneel onafhankelijk van PIP[4,5]2. Met dit voorbeeld is te zien dat een simpele transductie cascade vrij makkelijk gemoduleerd kan worden met een Bayesiaans netwerk. Wanneer een conditionele kansverdeling is gespecificeerd voor elk component kan er iets gezegd worden over de waarschijnlijkheid van actieve eiwitten.

In de volgende secties bespreek ik classificatie voor interacterende eiwitten. Ook zullen we zien hoe er een verzameling, van alle interacties in een bepaalde cel, kan worden geconstrueerd. Daarna kijken we naar een specifieke cascade en hoe deze gemodelleerd kan worden. Als laatst kijken we naar hoe er een graaf kan worden geconstrueerd van data uit een bepaalde cel en vergelijken we deze graaf met de consensus cascade uit de literatuur.

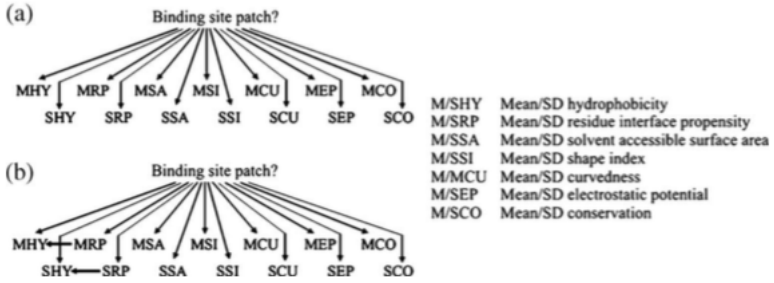
4.1 Eiwit-eiwit interactie Modelleren

Zoals gezegd in het vorige hoofdstuk zijn interacties de bouwstenen van een cascade. In de volgende secties bespreek ik Bayesiaanse classificatie voor eiwit−interacties.

4.1.1 Classificatie interactie

James R Bradford et al.[2] hebben Bayesiaanse netwerken gebruikt om eiwit oppervlaktes te classificeren als interacterend of niet−interacterend. De classificatie knoop classificeert als binding of geen binding. Een naïeve Bayesiaans netwerk en een Bayesiaans netwerk dat dit naïeve model uitbreidt met expert kennis, zijn vergeleken voor eiwit interactie classificatie. De netwerken hebben een *binding site patch* classificatie knoop. Deze knoop is de ouder van 14 andere knopen die het gemiddelde en standaardafwijking van 7 oppervlakte eigenschappen representeren. De eigenschappen zijn gebaseerd op expert kennis en de variabelen zijn continue. In Figuur 4.2 a is het naïeve Bayesiaanse netwerk weergegeven en in Figuur 4.2 b het Bayesiaanse netwerk weergegeven met 2 extra pijlen die gebaseerd zijn op expert kennis. De auteurs noemen dit netwerk daarom het expert netwerk.

Om de kans op interactie te bepalen werden er 180 eiwitten gebruikt. Voor elk van deze eiwitten werd er een oppervlakte gebied gekozen dat interacterend is en een gebied dat niet-interacterend is. Het Bayesiaanse netwerk berekent een kans op interactie voor een oppervlakte gebied. Classificatie is afhankelijk van een drempelwaarde. Wanneer de a-posteriori kans groter is dan de drempelwaarde dan classificeert het netwerk het gebied als interacterend. Wanneer de a-posteriori kans kleiner is dan de drempelwaarde classificeert het netwerk het ge-



Figuur 4.2: Bayesiaanse netwerk structuren voor classificatie van interacterende eiwitten. (a) Naïve Baysiaans netwerk. (b) Bayesiaans netwerk met 2 extra pijlen gebaseerd op expert kennis. *Insights into ProteinProtein Interfaces using a Bayesian Network Prediction Method* [2]

bied als niet−interacterend. Een ROC grafiek plot de sensitiviteit (true positive rate) tegen 1−specificiteit (false positive rate) over het domein [0.0,1.0]. Het gebied onder de ROC, *area under the curve* (AUC) geeft een maat voor de kwaliteit van de classificatie. 1.0 is een indicatie van een perfecte classificeerder en 0.5 is waardeloos omdat het dit gelijk staat aam willekeurigheid. De AUC van de naïve netwerk was 0.89±0.01 en die van de expert netwerk was 0.90±0.01. Omdat het verschil in AUC waarde niet groot is werd er gekozen om alleen verder te testen met het naïeve netwerk. Om kwaliteit van de voorspelling te beoordelen is de naive Bayesiaans netwerk getest met *leave one out cross validation*. Hierbij wordt het netwerk getraind met 179 eiwitten en daarna getest op 1 eiwit. Hierbij wordt gekeken of het eiwit correct wordt geclassificeerd. Dit werd gedaan voor elk eiwit zodanig dat elk eiwit een keer diende als test eiwit. Omdat niet-interacterende patches willekeurig gekozen werden is dit hele proces 5 keer opnieuw gedaan. Hiervan is het gemiddelde genomen. Er werd een gemiddelde succes score gemeten van 82 procent (148/180).

4.1.2 Probabilistische Interactoom

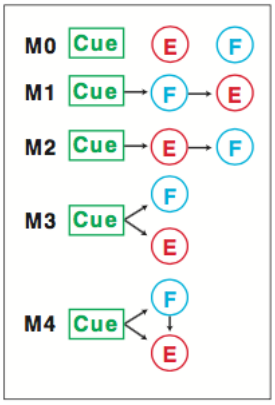
Een interactoom specificeert alle interacties in een bepaalde cel. Jansen et al [3] hebben onderzoek gedaan naar het interactoom in een gist cel. Zij hebben met behulp van Bayesiaanse netwerken een probabilistische interactoom (PI) geconstrueerd. Een probabilistisch interactoom is een netwerk waarin elk eiwit paar in een cel is geassocieerd met een kans die aangeeft of de eiwitten in hetzelfde complex zitten. Met een gouden standaard data set zijn twee verschillende Bayesiaanse netwerken getraind. De gouden data set bestaat uit geverifieerde classificatie data. De waarschijnlijkheid L(F) dat een eiwit paar in hetzelfde complex zit wordt gedefinieerd als de

fractie van goud standaard positieven die eigenschap f hebben gedeeld door de fractie negatieven met eigenschap f. Een eiwit paar werd als positief geclassificeerd als de gecombineerde waarschijnlijkheid groter is dan de drempelwaarde L_{cut}. Door combinatie van data uit 4 experimentele interactie datasets en een volledig verbonden Bayesiaans netwerk werd er een probalistische interactoom geconstrueerd: probabilistisch interactoom−experimenteel (PIE). Van genomische data en een naive Bayesiaans netwerk is er een probabilistisch interactoom−predicted geconstrueerd (PIP). De genomische data zijn gegevens verkregen uit erfelijke informatie. Om de netwerken te beoordelen is de data opgesplitst in een training en test gedeelte. Wanneer getest werd met de gouden standaard berekenden ze de fout positieven. Het bleek dat een drempelwaarde van 600 een bruikbare drempelwaarde is. Deze drempelwaarde is gebruikt voor verder testen. Het bleek dat de sensitiviteit van PIP ongeveer 27 procent is. Bij de PIE ligt dit percentage onder de 1 procent. Dit betekent dat er meer complexe interacties voorspelt kunnen worden met genomische data dan met experimentele data. De PIE en de PIP zijn gecombineerd tot een totale probabilistisch interactoom (PIT). Er is weinig informatie gegeven over de constructie van PIT. Omdat de PIE en PIP ongecorreleerd classificeren is voor de constructie van PIT een naïeve Bayesiaans netwerk gebruikt. Maar hoe het naïeve Bayesiaans netwerk precies is geïmplementeerd blijft onbekend.

4.1.3 MAPK cascade

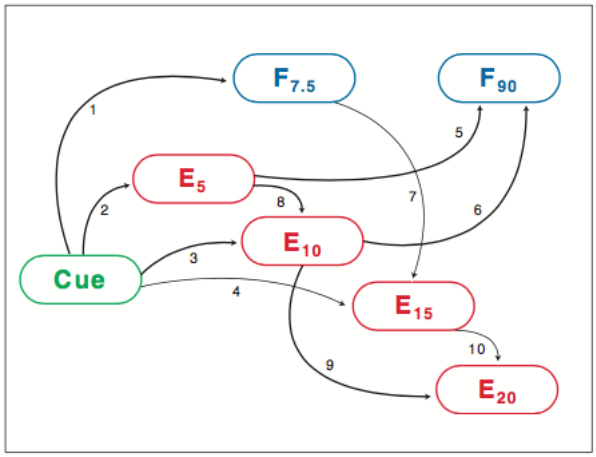
De mitogen-activated protein kinase (MAPK) route komt in elke cel voor en is belangrijk voor overleving, differentiatie en migratie. K. Sachs et al [8] hebben in deze cascade gekeken naar de activatie van focal adhesion kinase (FAK) en extracellulaire signaal-gereguleerde kinase (ERK), door interactie van het extracellulaire matrix eiwit, fibronectin (fn) en het transmembraan eiwit, integrin (a5b1). Vier Bayesiaanse netwerken zijn voorgesteld en afgebeeld in figuur 4.2. De Cue is de signalerende interactie tussen fn en a5b1. Modellen M1, M2, M3 en M4 zijn kandidaten voor model selectie en M0 is een controle model. In het controle model zijn er geen afhankelijkheden tussen de eiwitten. Model selectie is toegepast zoals besproken in hoofdstuk 2. Er zijn twee datasets gebruikt om de modellen te scoren. Dataset 1 bestond uit initile activatie van FAK en ERK. Dataset 2 bestond uit de globale activatie van FAK en ERK. Zoals verwacht scoorden M0 laag omdat all eiwitten onafhankelijk van elkaar waren. Data set 1 scoorde het best in combinatie met model M1: drie keer beter dan M2 en 4 keer beter dan M3 en M4. Dataset 2 scoorde het best met M4: 100 keer beter dan M1, 70 keer beter dan M2 en 10 keer beter dan

M3. Alhoewel, Model 1 en 4 het best uit de test komen is er geen significant verschil tussen de modellen. Volgens de auteurs zouden beslissende verschillen 1000 keer beter moeten zijn. Daarom concluderen ze dat er geen betrouwbare selectie gemaakt kan worden voor een van de voorgestelde modellen.



Figuur 4.3: MAPK cascade modellen, *Bayesian Network Approach to Cell Signaling Pathway Modeling* [8]

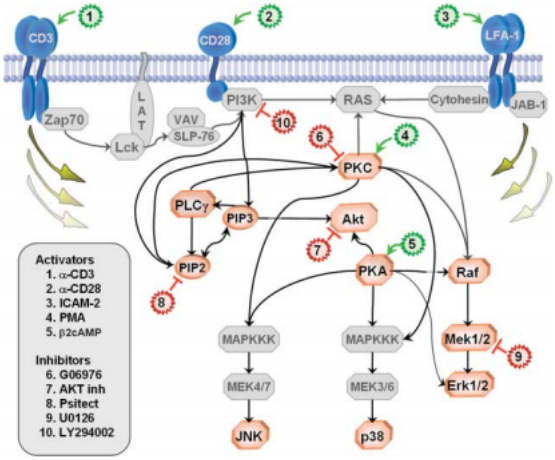
Om een structuur van een cascade te ontdekken kan er gebruik gemaakt worden van structuur leren zoals besproken in hoofdstuk 2. Het is computationeel onmogelijk om de hoogst scorende graaf te vinden omdat de zoekruimte te groot is. Maar er kan wel gekeken worden naar afhankelijkheden van hoog scorende modellen. Een dynamisch Bayesiaans netwerk is gebruikt om de MAPK structuur te zoeken met tijd−serie data. Tachtig datapunten voor ERK op 5,10,15 en 20 minuten en FAK op 7.5 en 90 minuten. In figuur 4.4 is de gevonden graaf weergegeven. Subscripts in het figuur staan voor tijd in minuten. De pijlen zijn genummerd voor referentie. Dunnere pijlen representeren de eigenschappen met een a−posteriori kans tussen de 0,5−0,85. Dikkere pijlen zijn consensus pijlen of hebben een a−posteriori kans van groter dan 0,85. Omdat pijlen geen kansen kunnen hebben is het onduidelijk wat er hier precies bedoeld wordt.



Figuur 4.4: Dynamische Baysiaans netwerk, *Bayesian Network Approach to Cell Signaling Pathway Modeling* [8]

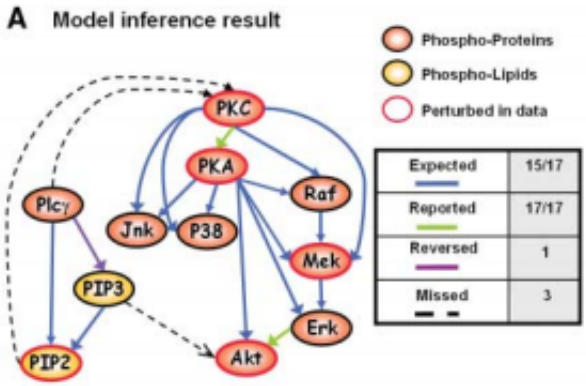
4.1.4 T-Cell Signaal netwerk

Naast de studie over MAPK cascade heeft K.Sachs et al.[9] ook nog de intracellulaire signaal cascade in een T cell (een afweercel) onderzocht onder invloed van de eiwitten CD3, CD28 en LFA−1. In de literatuur is er een geaccepteerde consensus over deze cascade en dit is weergegeven in Figuur 4.5



Figuur 4.5: MAPK cascade, *Causal Protein-Signaling Networks Derived from Multiparameter Single−Cell Data* [9]

Om een graaf voor deze T-cel te modelleren is er gebruik gemaakt van flow cytometry data [6] van 11 gefosforyleerde eiwitten en fosfolipides. PKC, PKA, Raf, Mek, Erk, Akt, Jnk, P38, PLcy, PIP3, PIP2. Elke onafhankelijke sample in de data set bestaat uit kwantitatieve hoeveelheden van elk van deze 11 gefosforyleerde moleculen, tegelijk gemeten in individuele cellen. De complete dataset werd geanalyseerd door een Bayesiaanse netwerk structuur inferentie algoritme. Het resulterende netwerk is weergegeven in figuur 4.6.



Figuur 4.6: Bayesiaans model voor MAPK cascade, *Causal Protein-Signaling Networks Derived from Multiparameter Single−Cell Data* [9]

Dit netwerk is vergeleken met de consensus uit de literatuur. Blauwe pijlen zijn de verwachte pijlen die bekend zijn. Groen pijlen zijn geen consensus pijlen maar wel minstens 1 keer aangetoond in de literatuur. Paarse pijlen zijn omgedraaid ten opzichte van de consensus. Ontbrekende pijlen zijn, ten opzichte van de consensus, aangeven met een stippellijn en behoren niet tot de gevonden graaf [9].

Hoofdstuk 5

Discussie en Conclusie

5.1 Bayesiaanse netwerken en signaal transductie

Eiwit interacties zijn een belangrijk onderdeel in de signaal transductie. Deze interacties zorgen voor conformatie veranderingen waardoor eiwitten actief of juist inactief worden. Er zijn verschillende Bayesiaanse netwerken gebruikt om deze interacties te voorspellen. Expert kennis over signaal transductie kan worden gebruikt om de stochastische variabelen te kiezen en netwerk structuren te bepalen. Data vanuit de signaal transductie kan gebruikt worden om de parameters en structuren te leren.

In de besproken literatuur is het niet altijd even duidelijk hoe Bayesiaanse netwerk structuren gekozen werden. Wat duidelijk naar voren komt is dat naïeve Bayesiaanse netwerken vaak worden gekozen. Dit komt omdat dit simpele netwerken zijn waar makkelijk mee geclassificeerd kan worden doordat ze geleerd kunnen worden uit data, de structuur vast staat en de kansen alleen geschat hoeven worden. Daarnaast geven de netwerken over het algemeen even goede resultaten vergeleken met complexere netwerken met extra afhankelijkheden. Hierbij moet wel in beschouwing worden gehouden dat dergelijke analyses afhankelijk zijn van de hoeveelheid beschikbare data. Dat wil zeggen dat wanneer er meer data beschikbaar zou zijn om te trainen, er eventueel een voorkeur zou kunnen ontstaan voor complexere netwerken die de afhankelijkheden in de werkelijkheid beter representeren [2].

Om eiwit interacties aan te tonen zijn er verschillende bio-moleculaire technieken benoemd in hoofdstuk 3. Deze technieken hebben een veel betere succes score dan de Bayesiaanse netwerken. Er kan met zekerheid vastgesteld worden dat bepaalde eiwitten daadwerkelijk met elkaar interacteren. Deze technieken zijn duur en kosten ook veel tijd. In tegenstelling tot Bayesiaanse

netwerken die in standaard software pakketten zitten en verder dan de ontwikkeling geen tijd en geld kosten.

5.2 Dynamische Bayesiaanse netwerken (DBNs)

In de signaal transductie bestaan er eiwitten die direct of indirect invloed hebben op zichzelf. In signaal transductie netwerken worden deze relaties geïnterpreteerd als feedback loops [1]. Het gebruik van een DBN zorgt ervoor dat er cyclische afhankelijkheden kunnen worden gerepresenteerd die een statisch netwerk niet kan representeren. Daarnaast kunnen in de DBN dynamische afhankelijkheden worden gerepresenteerd in de tijd [5]. Dit zijn twee belangrijke componenten in de signaal transductie. In de consensus modellen voor signaal transductie netwerken zijn er vaak feedback systemen weergegeven. Dergelijke weergaven zijn slechts statische weergaven van een feedback systeem [1]. Hoe de feedback systemen precies werken en gereguleerd worden blijft in zo'n model onduidelijk. Met een DBN wordt het netwerk complexer en geeft het meer informatie over hoe een feedback systeem werkt. Er kan statistisch bekeken worden hoe signaal transductie netwerken zich dynamisch ontwikkelen in een cel. Als mede bij het voorspellen van interacties moet hier ook in beschouwing worden genomen dat zulke netwerken erg gevoelig zijn voor de hoeveelheid data. Om iets realistisch te kunnen zeggen over signaal transductie netwerken gemoduleerd met een DBN moet er veel data beschikbaar zijn en het tijdsverloop moet expliciet in de data zitten[3][9].

5.3 Conclusie

- Wat voor soort Bayesiaanse netwerken kunnen worden gebruikt voor signaal transductie?

Naïeve Bayesiaanse netwerken, algemene Bayesiaanse netwerken en dynamische Bayesiaanse netwerken kunnen allemaal gebruikt worden voor onderzoek naar signaal transductie. De naïeve Bayesiaanse netwerken voor classificatie. De algemene Bayesiaanse netwerken kunnen als model staat voor simpelere consensus cascades. Dynamische Bayesiaanse netwerken kunnen de complexere cascades modelleren met tijd en feedback systemen.

- Wat voor soort signaal transductie data is er beschikbaar om te gebruiken?

Data voor onderzoek naar signaal transductie kan onder andere verkregen worden uit experimentele data van cellen, expert kennis uit de moleculaire biologie, eigenschappen van eiwitten

en genomische eigenschappen. Bayesiaanse netwerken zijn in staat om deze verschillende type van gegevens te combineren [3][4]. Maar ik denk dat het combineren van data pas echt tot zijn recht komt wanneer er veel data beschikbaar is. Een kritiekpunt op een Bayesiaans netwerk kan zijn dat er te weinig data beschikbaar is [8]. Ik stel dan ook voor om eerst meer data te hebben van dezelfde bron.

- Kunnen Bayesiaanse netwerken eiwit-interacties voorspellen? En hoe bruikbaar zijn deze voorspellingen?

Naïeve Bayesiaanse netwerken kunnen gebruikt worden voor eiwit interactie classificatie. Voor de keuze van de stochastische variabelen is denk ik expert kennis belangrijk. Dan is de toepassing naar mijn mening bruikbaar omdat er belangrijke eigenschappen voor interactie worden dan gebruikt. Wanneer er geen expert kennis wordt gebruikt kan er geclassificeerd worden op data die niets met de interacties te maken hebben.

- Kunnen Bayesiaanse netwerken cellulaire cascades modelleren? En hoe goed passen deze modellen op de consensus modellen?

Bayesiaanse netwerken kunnen signaal cascades modelleren. De conditionele probabilistische relaties tussen de componenten in signaal transductie zijn makkelijk te interpreteren in een BN. Het geleerde T-cel signaal netwerk, besproken in het vorige hoofdstuk, kan direct vergeleken worden met de consensus cascade [9]. Hierdoor denk ik dat de bevindingen makkelijk gebruikt kunnen worden in verder onderzoek naar bijvoorbeeld ziektes. Bayesiaanse netwerken beperken de cascade wel doordat ze geen tijd en feedback systemen kunnen representeren. Dynamische Bayesiaanse netwerken lossen dit probleem op. Mijn voorkeur voor het modelleren van signaal cascades gaat dan ook uit naar dynamische Bayesiaanse netwerken.

- Hoe kunnen Bayesiaanse netwerken toegepast worden bij onderzoek naar cellulaire signaal transductie en wat is de toegevoegde waarde?

De complexiteit van een signaal cascade is vaak onbekend. Een eiwit kan door vele andere eiwitten direct of indirect gereguleerd worden. Er bestaan interacties die relatief simpel zijn, maar ook invloed hebben in een complexere cascade [1]. Bayesiaanse netwerken kunnen gebruikt worden voor data analyse tussen variabelen in een gegeven dataset. De relaties tussen de variabelen zijn niet slechts lineair maar kunnen willekeurig complex zijn zodat er geen limiet is op het

aantal factoren dat betrokken is in de afhankelijkheid tussen de variabelen [8]. Door deze eigenschap kunnen ze goed gebruikt worden voor het moduleren van signaal transductie. Bayesiaanse netwerken hebben de potentie om zowel de simpelere interacties als de complexere cascades te modelleren. Omdat ik denk dat het toepassen van dynamische Bayesiaanse netwerken nieuwe inzichten kan geven voor de biomedische wetenschappen denk ik dat deze kunstmatige intelligentie tool een toegevoegde waarde kan hebben bij onderzoek naar cellulaire signaal transductie.

Bibliografie

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Molecular biology of the cell, Taylor and Francis Inc, zesde editie (2014).

[2] J.R. Bradford, C.J. Needham, A.J. Bulpitt and D.R. Westhead, Insights into ProteinProtein Interfaces using a Bayesian Network Prediction Method, Journal of Molecular Biology (2006).

[3] R. Jansen, H Yu, D Greenbaum, Y Kluger, N.J. Krogan, S. Chung, A Emili, M. Snyder, J.F. Greenblatt and M. Gerstein, A Bayesian Networks Approachfor Predicting Protein-Protein Interactions from Genomic Data, Science (2003).

[4] C.J Needham, J.R Bradford, A.J Bulpitt and David R Westhead, Inference in Bayesian networks, Nature Biotechnology (2006).

[5] C.J. Needham, J.R. Bradford, A.J. Bulpitt and D.R. Westhead, A Primer on Learning in Bayesian Networks for Computational Biology, PLoS Computational Biology (2007).

[6] D. Peer, Bayesian Network Analysis of Signaling Networks: A Primer, Science Signaling (2005).

[7] S.J. Russell and P Norvig, Artificial Intelligence: A Modern Approach, Pearson Education Limited, derde editie (2013).

[8] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D.A. Lauffenburger, Bayesian Network Approach to Cell Signaling Pathway Modeling, Science Signaling (2002).

[9] K. Sachs, O. Perez, D. Peer, D.A. Lauffenburger and G.P. Nolan, Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data, Science (2005).