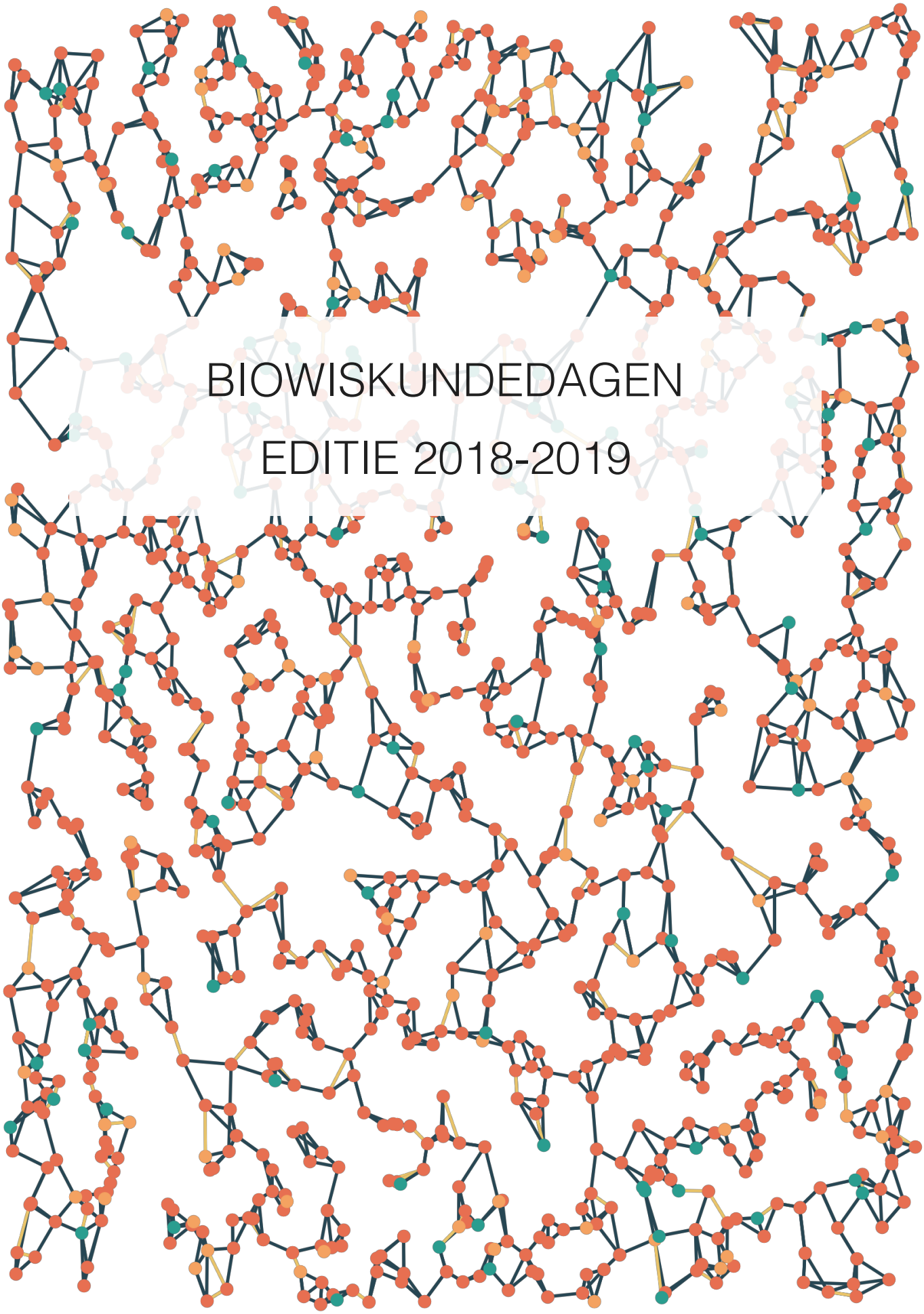


BIOWISKUNDEDAGEN

EDITIE 2018-2019

FACULTEIT
BIO-INGENIEURSWETENSCHAPPEN

Voorwoord

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent viverra convallis volutpat. Nunc sed est vel sem porta maximus. Nullam ut sagittis libero. Pellentesque fermentum finibus lacus eget tincidunt. Nulla rhoncus, enim ut tristique consequat, justo est pulvinar erat, id malesuada dolor orci non diam. Curabitur condimentum faucibus erat, in imperdiet dolor rhoncus at. Aliquam semper id tellus vel pharetra. Duis volutpat vitae orci sit amet scelerisque. Curabitur finibus ante non sollicitudin condimentum. Nunc maximus imperdiet faucibus.

Proin luctus gravida nunc, sit amet auctor quam malesuada eu. Phasellus vehicula commodo sollicitudin. In vitae nunc dictum, posuere ligula vel, aliquam lectus. Vivamus auctor tempus est, at accumsan leo bibendum eget. Duis tincidunt urna sed nibh posuere aliquet. Pellentesque tempus massa et nibh vehicula consequat. Nunc scelerisque, risus vel pretium varius, neque velit suscipit odio, et vehicula neque ex sit amet ex.

Suspendisse cursus dolor eros, non rhoncus ex gravida eget. Ut hendrerit, libero sit amet pretium accumsan, mauris leo aliquam sem, sit amet sagittis ipsum tellus nec est. Phasellus sapien lectus, volutpat placerat est in, aliquet pellentesque enim. Nam odio justo, feugiat a dapibus eu, interdum vel justo. Cras ligula eros, accumsan id cursus quis, blandit quis augue. Nullam volutpat sagittis tellus, id pretium elit porta eget. Donec rutrum urna turpis. Curabitur laoreet lorem vitae sapien ultrices finibus.

Cras quis neque sapien. Vestibulum at tincidunt neque. Vivamus tempor semper mattis. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Sed odio augue, molestie vitae quam eget, aliquam facilisis nibh. Maecenas gravida augue elementum sagittis vestibulum. Cras consequat tortor et mi suscipit, ac dictum nibh vehicula. Duis interdum ornare eros, ut iaculis mi maximus id. Integer blandit eget augue a finibus. Praesent mattis, ipsum ac tincidunt iaculis, sem ante consequat libero, pellentesque rutrum lorem ipsum quis lorem. Integer consectetur tincidunt ipsum id vestibulum. Cras vel rutrum tortor. Nam nec urna nisl. Ut elementum ultricies lectus et fringilla. Aliquam erat volutpat. Vestibulum pharetra massa eu interdum convallis.

Aenean eu libero at odio tempus suscipit. In tempus blandit mauris. Quisque vel vestibulum justo. Vestibulum mattis augue vel mauris porttitor euismod. Cras eget diam at elit vehicula consectetur. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec quis leo metus. Donec eu dapibus nulla. Aliquam venenatis lacinia dolor vel tristique. Sed ut mauris metus.

Januari 2019

Biowiskunde-team 2018-2019

- Ikke
- Een heleboel anderen

Inhoudsopgave

Eiwitten beter begrijpen met kansrekening

In dit project zullen we gebruik maken van kansrekening om de secundaire structuur (β-platen) van een eiwit te voorspellen. We overlopen eerst de basisregels van kansrekening en dan zullen we de kansformules vereenvoudigen om eenvoudiger te kunnen rekenen. Zo bekomen we een data-gedreven model om voor een aminozuur de kans te berekenen of dit in een β-plaat voorkomt of niet. Dit proces zullen we toepassen over volledige eiwitten via de glijdend venster methode. Ten slotte hebben we het kort even hebben over modelevaluatie: hoe betrouwbaar is zo'n model?

Een beetje achtergrond

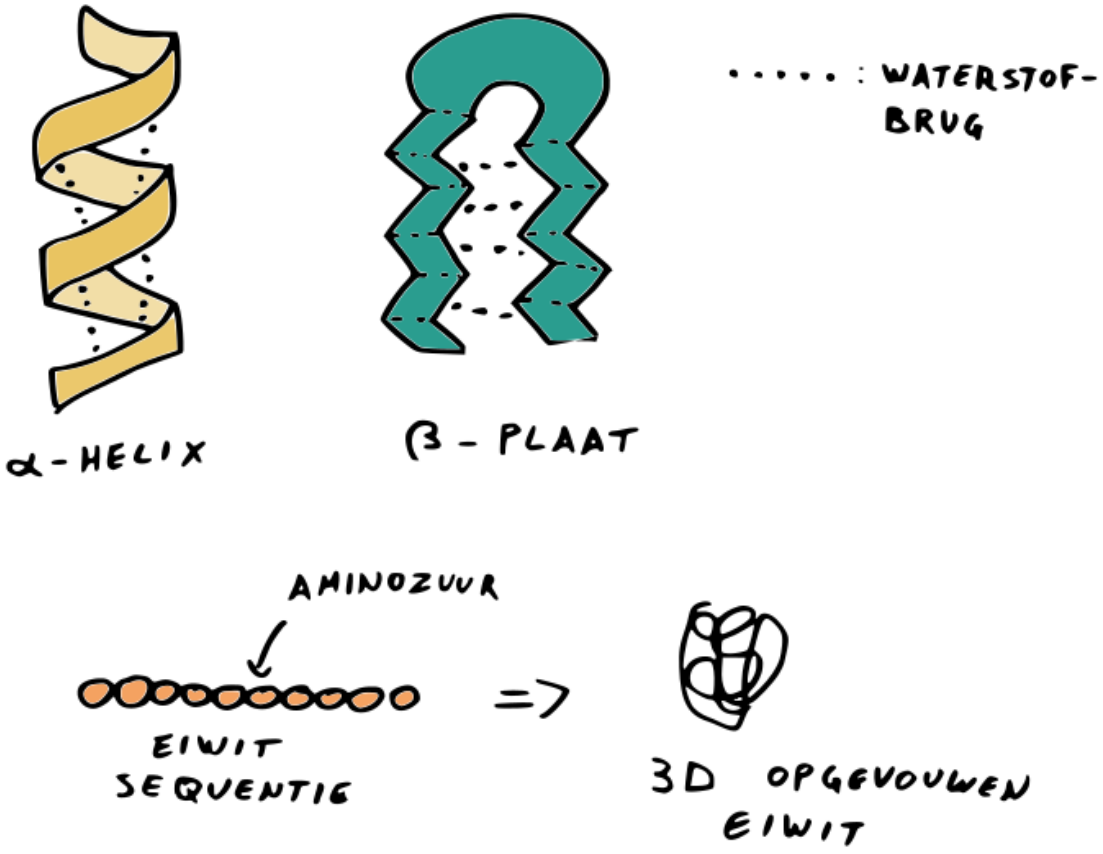
De wereld van eiwitten en hun opbouw

Eiwitten vormen één van de meest belangrijke klassen van biologische moleculen. Eiwitten vervullen actieve rollen en blijken essentieel te zijn voor zowat alle biologische processen in een levend organisme: ze staan in voor het verteren van voedsel, zorgen voor ontgifting, geven informatie door, maken beweging mogelijk en nog veel meer. Zo vormen ze een fundamenteel onderdeel van elk biologisch wezen.

De opbouw van een eiwit is relatief simpel. Net zoals DNA is een eiwit een *polymeer*: een lange streng van meer eenvoudige moleculen. Deze eenvoudige moleculen worden *aminozuren* (AZ) genoemd, waarvan er in de natuur 20 verschillenden voorkomen, elk voorgesteld door een hoofdletter. Een specifieke opeenvolging van dergelijke aminozuren wordt de *primaire structuur* van een eiwit genoemd. Deze primaire structuur bepaalt hoe een eiwit zich verder zal opvouwen (in een functionele 3D-structuur) en legt de biologische functie van een eiwit vast.

Uit de primaire structuur van eiwitten volgt de *secundaire structuur*. Deze structuren onstaan door waterstofbruggen (niet-covalente bindingen tussen vrije waterstoffen en hydroxylgroepen) tussen naburige aminozuren (zie Figuur 1). De belangrijkste secundaire structuren zijn α-*helices* en β-*platen* (Engels: β-sheets). Gegeven dat deze secundaire structuur enkel door de primaire structuur bepaald wordt, kunnen we wiskunde gebruiken om de secundaire structuur te voorspellen[1].

[1] Naast de primaire en secundaire structuur hebben eiwitten doorgaans ook en *tertiaire* en *quaternaire* structuur. De tertiaire structuur is de globale opvouwing van het eiwit en is veel (veeeeeeeel) moeilijker om computationeel te bepalen. De quaternaire structuur omvat hoe verschillende eiwitten samen een groter complex vormen.

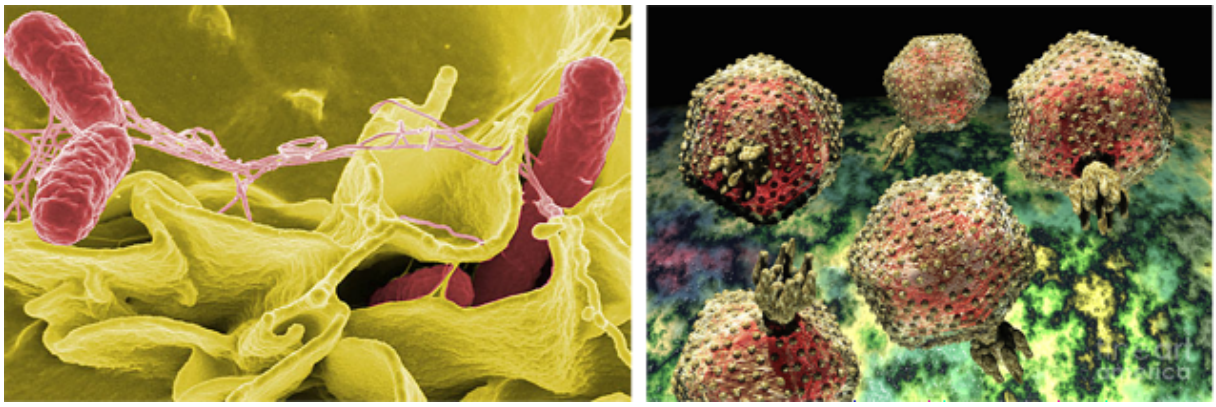


Figuur 1: (boven) Eiwitsequenties vormen secundaire structuren via niet-covalente waterstofbindingen. (onder) Een eiwit is slechts een aaneenschakeling van aminozuren die zich zelfstandig opvouwen tot een driedimensionale structuur.

Stoute bacteriën en goede virussen

Zelfs de allerkleinste biologische entiteiten, de virussen, gebruiken eiwitten voor infectie, wat nodig is voor hun vermenigvuldiging. Een zeer interessante groep virussen zijn de bacteriofagen of kortweg fagen. Dit zijn virussen die bacteriën infecteren en ook kunnen afdoden. Bacteriën, en dus ook fagen, komen overvloedig voor in ons lichaam. Vele bacteriën zijn goedaardig en helpen ons lichaam optimaal functioneren. Soms dringen echter pathogene bacteriën ons lichaam binnen en maken ze ons

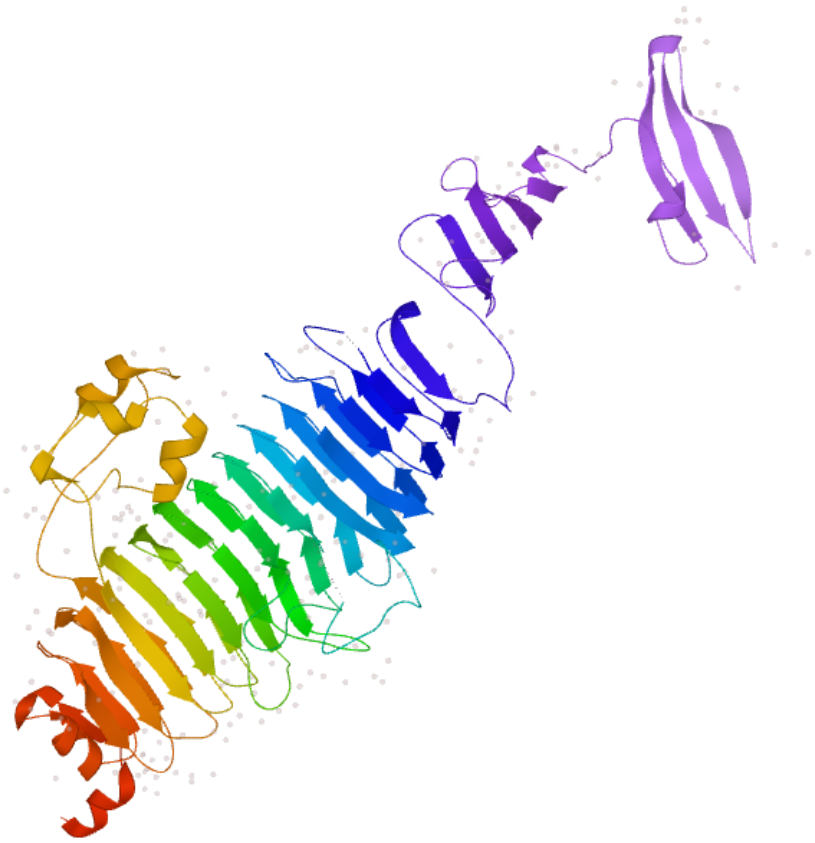
ziek. *Salmonella enterica* is zo'n bacterie (Figuur 2). *Salmonella* dringt ons lichaam binnen via besmet voedsel: de bacterie kan overleven op onvoldoende verhitte eieren en vlees, alsook op rauwe groenten en fruit. Eens de bacterie zich in onze darmen bevindt, kan ze ons ernstig ziek maken.



Figuur 2: (links) Microscopische figuur van de *Salmonellabacterie*. (rechts) Figuur van P22 fagen die *Salmonella* kunnen infecteren.

Een bijkomend probleem is dat *Salmonella* en andere bacteriën steeds meer resistent worden tegen antibiotica. Gelukkig kunnen we ook fagen inzetten om bacteriën te bestrijden! Cruciaal voor een faag bij het infecteren van zijn bacteriële gastheer zijn specifieke faag eiwitten die componenten van de salmonellabacterie herkennen. Deze faag eiwitten verschillen vaak tussen verschillende salmonellafagen. Hierdoor kunnen verschillende fagen andere varianten van de *Salmonella* bacterie herkennen. Anderzijds komt er tussen verschillende salmonellafagen ook vaak een geconserveerd eiwitdomein voor. Dit is een stukje van het eiwit dat wel hetzelfde is tussen de verschillende salmonellafagen. Bij salmonellafagen is dit een zogenaamd β-helicaal domein. Dit domein vormt als het ware een *moleculaire boor* die de celwand van de bacterie kan doorboren, wat nodig is om de infectie te starten. Door zo'n faag eiwitten beter te begrijpen kunnen we ze daarna ook beter inzetten tegen gevaarlijke bacteriën.

Een voorbeeld van zo'n faag eiwit is het staarteiwit van Salmonellafaag P22: *P12528* (zie rechtse deel van Figuur 2). Tussen aminozuur 140 en 543 bevindt zich een groot β-helicaal domein (bestaande uit parallele β-platen) dat een puntig einde heeft rond aminozuur 113 (Figuur 3). De aanwezigheid van die β-platen is belangrijk voor de specifieke functie van het eiwit. Deze secundaire structuren (de β-platen) kunnen we bestuderen via wiskunde en computers. Dit onderzoeksdomein noemen we *bio-informatica*. In bio-informatica wordt wiskunde gecombineerd met computerkracht om interessante biologische fenomenen te bestuderen en biologische problemen op te lossen.



Figuur 3: Het P12528 eiwit, ook wel Salmonellafaag P22 *tail spike* eiwit genoemd. Dit eiwit bestaat uit een uitzonderlijk groot aantal β-platen die samen een complexe boorkop vormen. Regenboogkleuring in volgorde van de sequentie.

In dit project zetten we de computer aan het werk om eiwitten te bestuderen. Zo'n eiwitten bestuderen wetenschappers vaak op basis van de aminozuursequentie van het eiwit. Door specifieke instructies te geven aan de computer kunnen we voorspellingen maken voor β-platen om zo de β-helicale domeinen te vinden! In dit project zullen we de computer leren om dergelijke voorspellingen te maken. Hieronder bekijken we eerst welke wiskunde je daar net voor nodig hebt.

Rekenen met kansen en de regel van Bayes

Kansrekening in een notendop

Kansrekening of probabiliteitstheorie is de tak van de wiskunde die zich bezig houdt met *kansen*. Met kansen kom je dagelijks in contact, denk maar aan gezelschapsspellen waarbij je moet dobbelen of Frank Deboosere die aangeeft dat er 60%[2] kans op neerslag is voor morgen. Er zijn nog vele andere voorbeelden, en net omdat kansrekenen zo belangrijk is in het dagelijkse leven, is het interessant om dit te bestuderen.

Er zijn enkele fundamentele regels die steeds gelden bij het berekenen van kansen. Figuur 4 stelt deze regels visueel voor:

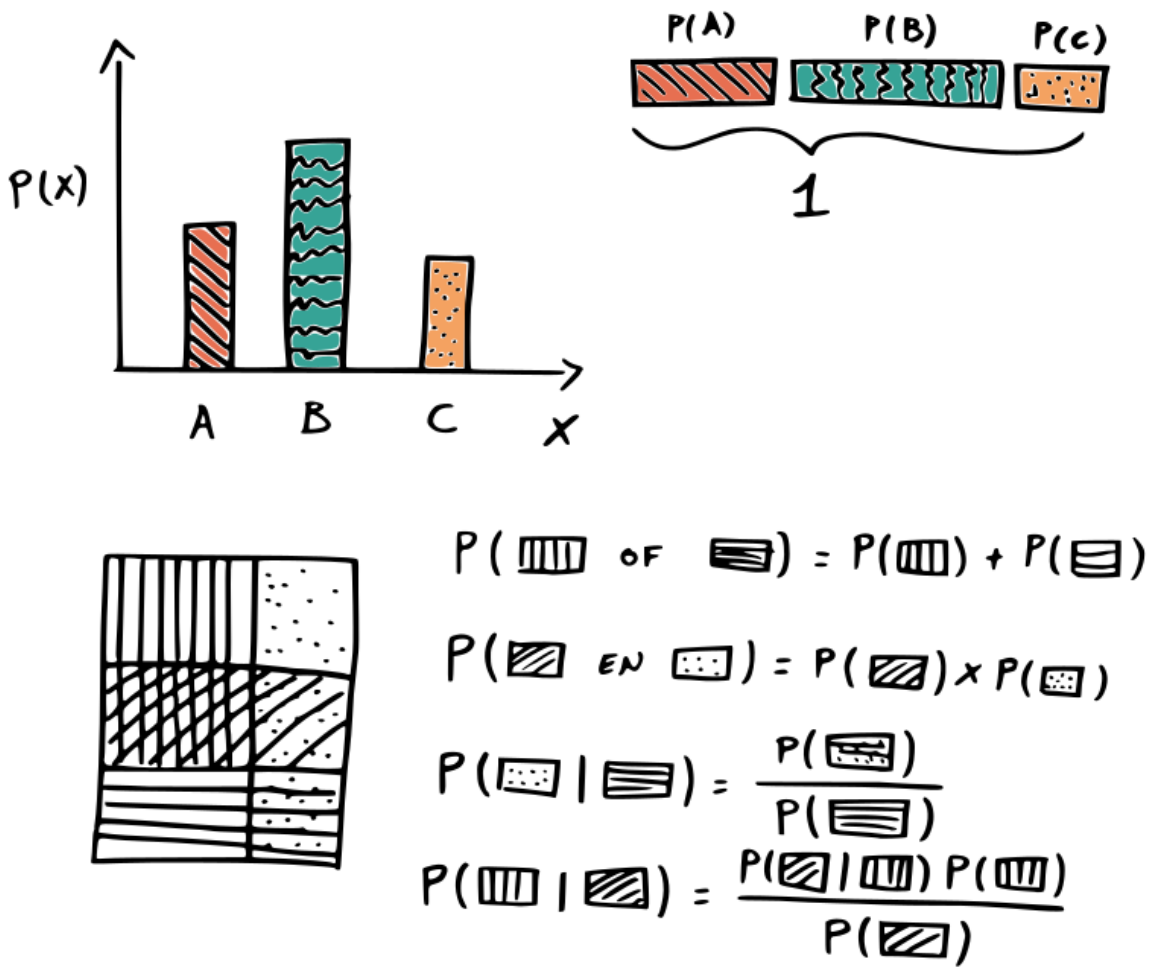
1. De kans van een gebeurtenis is een niet-negatief getal (nul inclusief) [3].
2. De totale kans dat er *een* gebeurtenis plaatsvindt is 1 (honderd procent)[4].
3. (**somregel**) De kans dat één van twee elkaar uitsluitendende gebeurtenissen plaatsvindt is de som van de kansen van die gebeurtenissen[5].
4. (**productregel**) Bij twee *onafhankelijke* gebeurtenissen is de kans dat beide gebeurtenissen samen plaatsvinden het product van die kansen[6].
5. Er bestaan *conditionele* kansen, dit is de kans dat een gebeurtenis A plaatsvindt gegeven een gebeurtenis B. De conditionele kans[7] wordt berekend als:

$$P(A \mid B) = \frac{P(A \text{ en } B)}{P(B)}\,.$$

6. Met de *regel van Bayes* kunnen we via de kansen van een gebeurtenis A de kansen berekenen voor een andere gebeurtenis B. De formule wordt hieronder gegeven voor gebeurtenissen A en B:

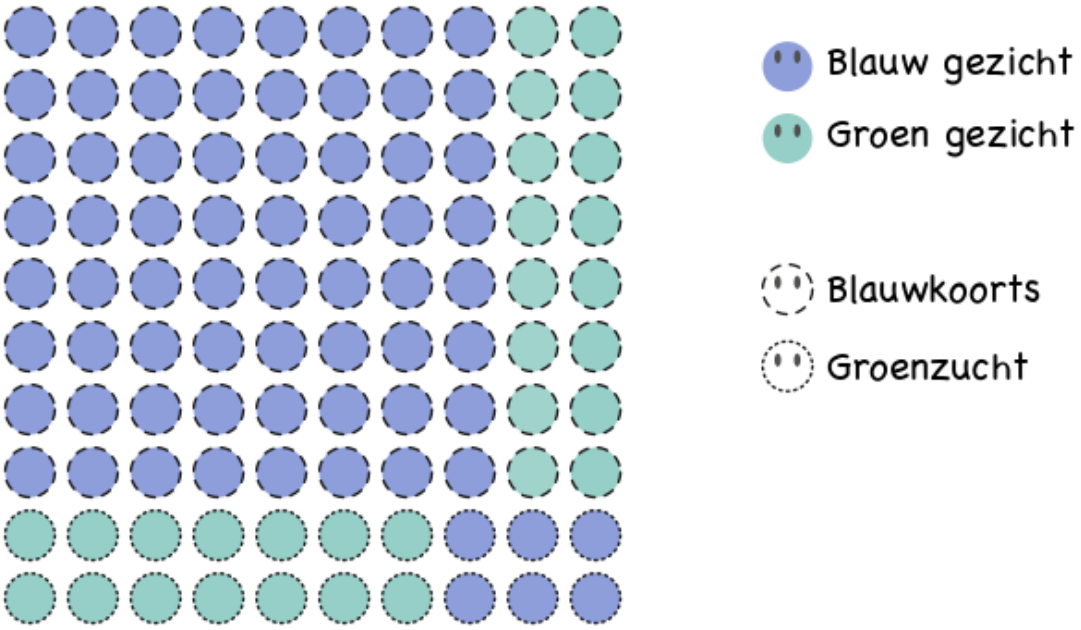
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}\,.$$

[2] Vergeet niet bij het rekenen dat 60% = 0.60.
[3] Bijvoorbeeld, de kans op een 6 gooien met een zesogige dobbelsteen is 1/6.
[4] De som van de kansen van alle mogelijke uitkomsten van een worp van een dobbelsteen is 1/6+1/6+…+1/6=1.
[5] Bijvoorbeeld, met een dobbelsteen gooien kan je nooit én een even getal gooien (kans van 3/6=1/2) én een drie gooien (kans van 1/6). De kans op één van beide gebeurtenissen is 3/6+1/6=4/6=2/3.
[6] Bijvoorbeeld, de kans dat je bij twee opeenvolgende worpen van een dobbelsteen twee keer een zes gooit is $1/6 \cdot 1/6 = 1/36$.
[7] Bijvoorbeeld, de kans dat we met een dobbelsteen een zes gooien gegeven dat het een even getal was is $(1/6)/(1/2) = 1/3$.



Figuur 4: Voorstelling van de basisregels van kansrekening. Kansen voor beurtenissen worden voorsteld door niet-negatieve waarden die samen tot 1 sommeren.

> **Oefening 1:** Je wordt op een nacht rillend van de koorts wakker. Je moet overgeven en hebt overal jeuk. Er zijn twee ziekten met deze symptomen: blauwkoorts en groenzucht. De ene ziekte komt vaker voor dan de andere: wie ziek is heeft in 80% van de gevallen last van blauwkoorts, terwijl groenzucht slechts in 20% van de gevallen voorkomt. Zoals de naam doet vermoeden, hebben deze ziekten nog een ander, duidelijk zichtbaar symptoom. Mensen met blauwkoorts krijgen doorgaans een blauw gezicht en deze met groenzucht een groen gezicht. **In 20% van de gevallen krijgt een persoon met blauwkoorts een groen gezicht en in 30% van de gevallen krijgt iemand met groenkoorts een blauw gezicht!** Je spoedt je naar de spiegel en iemand met een groen gezicht staart terug. Welke ziekte heb je meest waarschijnlijk? Bekijk onderstaande figuur en vul de ontbrekende kansen in de tabel verder aan.



Figuur 5: Honderd individuen met blauwkoorts en groenzucht. Sommigen hebben een blauwe gezichtskleur, anderen een groene.

$$P(\text{blauwkoorts}) = ... \qquad P(\text{groenzucht}) = ...$$

ziekte	kans blauw gezicht	kans groen gezicht
blauwkoorts	…	…
groenzucht	…	…

$$P(\text{blauw gezicht}) = ... \qquad P(\text{groen gezicht}) = ...$$

$$P(\text{blauwkoorts} \mid \text{groen gezicht}) = ... \qquad P(\text{groenzucht} \mid \text{groen gezicht}) = ...$$

Naive Bayes

Nu we de regel van Bayes intuïtief begrijpen, kunnen we deze toepassen voor het voorspellen van β-platen in eiwitten. De *naive Bayes*-methode kan hiervoor gebruikt worden. Deze methode maakt gebruik van de regel van Bayes om voorspellingen te maken o.b.v. een gegeven input. In dit project willen we een β-plaat voorspellen o.b.v. de eiwitsequentie (de input).

De regel van Bayes kan voor dit geval als volgt geschreven worden:

$$P(\beta\text{-plaat} \mid \text{eiwitsequentie}) = \frac{P(\text{eiwitsequentie} \mid \beta\text{-plaat})P(\beta\text{-plaat})}{P(\text{eiwitsequentie})}$$

De eerste kans in de breuk kunnen we wat herschrijven als:

$$P(\text{eiwitsequentie} \mid \beta\text{-plaat}) = P(A_1 A_2 \ldots A_n \mid \beta\text{-plaat})\,.$$

Hierboven schrijven we dus de eiwitsequentie gewoon als de opeenvolging van de n aminozuren. Hier stelt A_i de identiteit voor van het aminozuur op positie i. Hoe berekenen we de kans op een gegeven sequentie? Voor een stukje met lengte $n = 10$ hebben we $20^{10} = 10240000000000 \approx 10^{13}$ unieke sequenties. In een probabilistisch model moeten we dus een vereenvoudiging doorvoeren!

> **Vereenvoudiging:** We gaan ervan uit dat de kans dat een bepaald aminozuur op een bepaalde plaats voorkomt **onafhankelijk is van de aminozuren op elke andere plaats** (ergens huilt er een moleculair bioloog).

Deze vereenvoudiging is natuurlijk volstrekt biologisch onrealistisch! De aminozuren zijn in werkelijkheid juist erg afhankelijk, bijvoorbeeld omdat twee aminozuurtjes met elkaar in contact komen en een waterstofbrug vormen. Hoe fout deze vereenvoudiging ook is, ze is echter wel nuttig!

Het laat ons toe om de kansen voor een β-plaat redelijk goed te benaderen! In formulevorm is deze veronderstelling voor ons probleem[8][9]:

$$P(\text{eiwitsequentie} \mid \beta\text{-plaat}) \approx P(A_1 \mid \beta\text{-plaat})P(A_2 \mid \beta\text{-plaat}) \dots P(A_n \mid \beta\text{-plaat})$$
$$= \prod_{i=1}^{n} P(A_i \mid \beta\text{-plaat})$$

Uiteindelijk kunnen we de regel van Bayes dus als volgt noteren om β-platen te voorspellen:

$$P(\beta\text{-plaat} \mid \text{eiwitsequentie}) \approx P(\beta\text{-plaat}) \prod_{i=1}^{n} \frac{P(A_i \mid \beta\text{-plaat})}{P(A_i)} .$$

We kunnen de kans op een β-plaat gegeven een sequentie dus berekenen aan de hand van termen die we makkelijk uit data kunnen schatten door te tellen! De kans dat het aminozuur voorkomt in een β-plaat gedeeld door de kans dat dat aminozuur in een willekeurig eiwit voorkomt wordt de **odds** (Nederlands: *kansverhouding*) genoemd. Hoewel we dit hier niet zullen doen, is het misschien ook wel belangrijk te vermelden dat in de praktijk de logaritme[10] van deze kansen genomen wordt om de berekeningen te vereenvoudigen.

> **Vraag:** Wanneer zou je stellen dat een regio waarschijnlijk een β-plaat is?

Een eenvoudig voorbeeld met de hand uitwerken

Nu kunnen we de bovenstaande formule gebruiken om voorspellingen te maken voor eiwitten. Vooraleer we deze formule doorgeven aan de computer zullen we ze eerst zelf op papier eens toepassen. Hiervoor hebben we volgende kansen nodig (rechterlid van de bovenstaande formule):

- $P(A_i)$: de probabiliteiten van het voorkomen van elk aminozuur A_i.
- $P(\beta\text{-plaat})$: de kans om een β-plaat waar te nemen.
- $P(A_i \mid \beta\text{-plaat})$: de kans om een bepaald aminozuur waar te nemen, gegeven dat de sequentie een β-plaat is.

[8] On helemaal precies te zijn, de veronderstelling zegt dat $P(A_1 A_2 \dots A_n) \approx P(A_1)P(A_2) \dots P(A_n)$, wat zegt dat aminozuren onafhankelijk zijn. Wij gebruiken de identiteit $P(A_1 A_2 \dots A_n \mid \beta\text{-plaat}) \approx P(A_1 \mid \beta\text{-plaat})P(A_2 \mid \beta\text{-plaat}) \dots P(A_n \mid \beta\text{-plaat})$, wat zegt dat aminozuren onafhankelijk zijn binnen een β-plaat. Deze twee uitspraken zijn **niet** inwisselbaar!

[9] Hier maken we gebruik van de notatie voor een product: $\prod_{i=1}^{n} x_i = x_1 x_2 \dots x_n$, (bv. $\prod_{i=2}^{4} i = 2 \times 3 \times 4 = 24$).

[10] De logaritme met basis 10 wordt gedefineerd als:

$$\log_{10} x = y \iff 10^y = x .$$

Herinner je dat voor positieve getallen a en b geldt dat $\log(ab) = \log(a) + \log(b)$ en $\log(a/b) = \log(a) - \log(b)$. Wetenschappers gebruiken logaritmen vaak om vermenigvuldigingen in sommen om te zetten. Het voordeel met logaritmen is dat heel kleine getallen door vermenigvuldiging in negatieve waarden omgezet worden.

Oefening 2: Onderstaande tabel bevat experimenteel bepaalde aminozuur (AZ) aantallen van een staarteiwit van een faag die we zullen gebruiken om voorspellingen te maken. Op basis van de aantallen en het totaal aantal aminozuren kan je de ontbrekende kansen in de tabel berekenen, alsook de kans op een β-plaat. Deze kansen heb je nodig om de formule uit te werken. Vervolledig deze tabel.

AZ	totaal aantal	$\mathbf{P}(\mathbf{A_i})$	aantal in β-plaat	$\mathbf{P}(A_i \mid \beta\text{-plaat})$	$\frac{\mathbf{P}(A_i \mid \beta\text{-plaat})}{P(A_i)}$
A	48	…	21	…	…
C	8	0.0120	2	0.0060	0.5038
D	48	…	19	…	…
E	22	…	11	…	…
F	25	0.0375	13	0.0393	1.0479
G	71	0.1064	29	0.0876	0.8231
H	10	0.0150	4	0.0121	0.8060
I	51	…	36	…	…
K	34	…	12	…	…
L	49	0.0735	30	0.0906	1.2337
M	9	0.0135	6	0.0181	1.3434
N	41	0.0615	18	0.0544	0.8847
P	28	0.0420	7	0.0211	0.5038
Q	22	0.0330	9	0.0272	0.8244
R	23	0.0345	14	0.0423	1.2266
S	50	…	25	…	…
T	46	0.0690	23	0.0695	1.0076
V	48	0.0720	31	0.0937	1.3014
W	7	0.0105	2	0.0060	0.5757
Y	27	…	19	…	…
Totaal	**667**	-	**331**	-	-

Dus is de kans op een β-plaat gelijk aan:

$$P(\beta\text{-plaat}) = ...$$

> **Oefening 3:** Volgende korte sequentie is een klein deeltje van het P22 staarteiwit: „YSIEADKK". Experimenteel werd reeds bepaald dat dit geen β-plaat is, maar een α-helix. Bereken nu via de laatst geziene formule de kans dat die sequentie een β-plaat bevat (deze kans zou klein moeten zijn). Maak gebruik van de tabel met kansen die je net hebt ingevuld.

i	A_i	$\frac{P(A_i \mid \beta\text{-plaat})}{P(A_i)}$
1	…	…
2	…	…
3	…	…
4	…	…
5	…	…
6	…	…
7	…	…
8	…	…

Dus is de conditionele kans op een β-plaat gegeven de eiwitsequentie bij benadering gelijk aan:

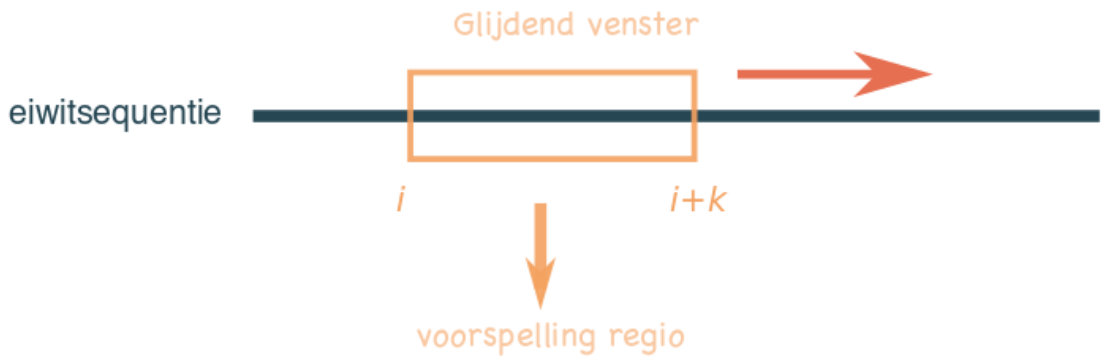
$$P(\beta\text{-plaat} \mid \text{eiwitsequentie}) \approx ...$$

Naive Bayes op de computer

Glijdende vensters en drempelwaarden

In het computerdeel van dit practicum gaan we nu de Naive Bayes-methode toepassen op het volledige P22 eiwit dat we eerder besproken hebben. Het doel is om te ontdekken waar de β-platen zich in het eiwit bevinden. We zullen β-platen voorspellen met behulp van de Naive Bayes-methode en de voorspellingen (i.e. de kansen) dan voorstellen via een grafiek. Hiervoor bewegen we aminozuur voor aminozuur over het eiwit via een *glijdend venster* van lengte k. In dit glijdend venster kijken we naar de aminozuren op elke positie van i tot $i + k$ en vermenigvuldigen alle odds voor elk aminzozuur in dit venster (i tot $i + k$). We noteren dit als

$$s_i^k = P(\beta\text{-plaat}) \prod_{j=i}^{i+k} \frac{P(A_j \mid \beta\text{-plaat})}{P(A_j)} .$$



Figuur 6: Illustratie van het glijdend venster over een sequentie.

In elke stap (voor elk glijdend venster) maken we een voorspelling die we later visueel kunnen voorstellen in een plot.

Eerder hebben we gesteld dat we een regio als een β-plaat classificeren indien

$$P(\beta\text{-plaat} \mid \text{eiwitsequentie}) > 0.5 .$$

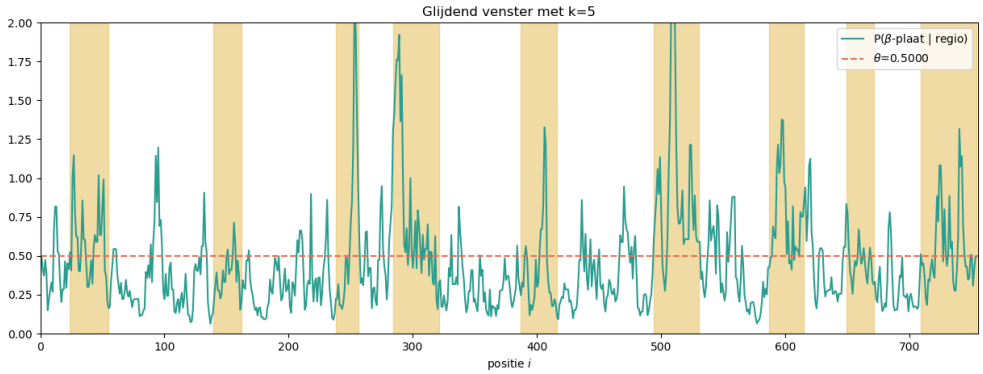
We willen dit echter veralgemenen zodat we strenger of minder streng kunnen zijn om secundaire structuren te vinden:

$$P(\beta\text{-plaat} \mid \text{eiwitsequentie}) > \theta .$$

Hier is θ een zorgvuldig gekozen *drempelwaarde* (Engels: threshold). De keuze van θ heeft gevolgen voor de correctheid van onze voorspellingen:

- als we θ te hoog kiezen is onze drempelwaarde te streng en zullen we dus bepaalde regio's niet als β-platen voorspellen terwijl dit eigenlijk wel β-platen zijn.
- als we θ te laag kiezen zijn we niet streng genoeg. We zullen dus regio's voorspellen als β-plaat dat eigenlijk geen β-plaat zijn.

Hieronder zie je een voorbeeld van een analyse met een glijdend venster.



Figuur 7: Voorstelling van de resultaten van het glijdend venster. De regio's waar de conditionele kans de drempelwaarde overschreid gedurende een bepaalde lengte (bepaald door k), worden aangeduid als β-plaat.

Je kan dus inzien dat we de waarde van θ net goed willen kiezen zodat we het aantal foute voorspellingen tot een minimum beperken. Dit beperken van foute voorspellingen is altijd gewenst bij het gebruik van wiskundige modellen, en om deze fouten te bestuderen doen we aan *modelevaluatie*.

Modelevaluatie: op welke manier is je model fout?

Wiskundige modellen maken zelden perfecte voorspellingen. Toch is het in de praktijk belangrijk dat modellen zeer accurate voorspellingen maken. Als bijvoorbeeld een zelfrijdende auto een foute voorspelling maakt over waar hij moet rijden kan dat mogelijks fataal zijn voor personen in de wagen en/of in de omgeving. Wanneer een wiskundig model voorspelt dat jij een kankergezwel hebt terwijl dat eigenlijk niet zo is, krijg je onnodig dure chemotherapie (die vaak ook slechte bijwerkingen heeft). Er zijn natuurlijk ook minder ernstige voorbeelden: wanneer het algoritme van Netflix je weer een serie aanraadt die je niet goed vindt, ga je simpelweg niet naar die serie beginnen kijken. Maar uiteraard wil ook Netflix zijn klanten de meest relevante films en series aanraden, en dat doen ze door continu voorspellingen te maken o.b.v. de series en films die jij al bekeken hebt en de grote hoeveelheid data die ze over hun andere klanten hebben.

Om inzicht te krijgen in hoe goed of hoe slecht een model voorspellingen maakt, zullen we het model evalueren: we bepalen hoe goed het model werkt op nieuwe data. Bij het voorspellen van secundaire structuren kan ons model slechts twee soorten voorspellingen maken: ofwel is de beschouwde regio onderdeel van een β-plaat ofwel is die dat niet. Het eerste noemen we een *positieve voorspelling*, het tweede een *negatieve voorspelling*. Deze terminologie is afkomstig uit de geneeskunde: een diagnostische test is positief als de persoon ziek is, en negatief als de persoon niet ziek is. In onze context

hebben we echter geen voorkeur voor een positieve of negatieve voorspelling, we willen enkel correcte voorspellingen! Ons model kan twee soorten foute voorspellingen maken:

- Er werd foutief voorspeld dat een regio deel uitmaakt van een β-plaat. Dit heet een **vals positieve** voorspelling (Engels: *false positive*).
- Een regio werd voorspeld als geen deel van een β-plaat terwijl dit in werkelijkheid wel zo is. Dit is een **vals negatieve** voorspelling (Engels: *false negative*).

De correcte en foute voorspellingen kunnen we eenvoudig voorstellen in een tabel:

	Voorspeld als β-plaat	**Voorspeld als geen β-plaat**
Regio is deel van β-plaat	echt positief	vals negatief
Regio is geen deel van β-plaat	vals positief	echt negatief

Beide foute voorspellingen zijn nauw verbonden met de keuze van de drempelwaarde θ, alsook de grootte van het glijdend venster. In een laatste stap zullen we daarom de drempelwaarde θ en de grootte van het glijdend venster manueel aanpassen en het effect bestuderen op het aantal foute voorspellingen. Op die manier kunnen we θ en de grootte van het venster optimaal kiezen en zo de foute voorspellingen tot een minimum beperken.

Stappenplan

Concreet zullen we de computer dus instructies geven om het volgende te doen:

1. Startend bij het begin van een eiwitsequentie maakt de computer een eerste voorspelling voor het stukje van de sequentie dat zich in het glijdend venster bevindt. Dit doet hij door Naive Bayes toe te passen en het stukje sequentie als β-plaat te voorspellen wanneer de berekende kans groter is dan de vooropgestelde drempelwaarde θ.
2. Daarna schuift de computer het glijdend venster één aminozuur op in de sequentie en maakt een nieuwe voorspelling voor dit glijdend venster. Dit proces herhaalt de computer tot het einde van de eiwitsequentie bereikt is.
3. Voor elk glijdend venster slaat de computer de voorspelling op, zodat die later visueel voorgesteld kan worden.
4. We laten de computer de voorspellingen vergelijken met de werkelijke secundaire structuren, zodat we het model kunnen evalueren o.b.v. vals positieven en vals negatieven.
5. Als laatste stap veranderen we manueel de drempelwaarde θ en de grootte van het glijdend venster, om op die manier te proberen de vals positieven en vals negatieven tot een minimum te beperken.

> **Computeroefeningen:** Het glijdend venster om secundaire structuur te voorspellen is beschikbaar in een Jupyter notebook. Deze zijn beschikbaar via de website van de biowiskundedagen (http://www.biowiskundedagen.ugent.be/) of door op *deze link* te klikken. Met die notebook kan je experimenteren met de grootte van het venster, de parameter k en de drempelwaarde. Enkele vragen hierbij zijn de volgende: Welke invloed heeft de parameter k als die groter wordt? Wat wil $k = 1$ zeggen? Welke invloed heeft het verhogen en verlagen van de drempelwaarde op de verschillende types fouten? Kan je het aantal valse positieven (niet-β-platen die als β-plaat voorspeld worden) zo laag mogelijk krijgen? Hoe zorg je er voor dat je geen enkele β-plaat mist? Wat is het nadeel hiervan?

Naive Bayes programmeren

We illustreren hier hoe de kansen op de computer berekend kunnen worden in de programmeertaal Python.

```
# volledige eiwitsequentie
P22eiwit = "MTDITANVVVSNPRPIFTESRSFKAVANGKIYIGQIDTDPVNPANQIPVYIENEDGSHVQITQPLIIN

# voorbeelden van sequenties van beta-platen
beta_platen = ["SRSF", "KIYIGQ", "VYIE", "HVQI", "QPLII", "IVY", "IVTVQ",
               "SMAIY", "QVDYIA", "SVK", "YPT", "VDGLLI", "TVD", "TIEC",
               "AKFI", "DGNLIFT", "RIAG", "FME", "WVI", "KTDGY", "STLEIRE",
               "EVHR", "SGLMAGFLFRG", "KMVD", "NNPSG", "IITFE", "LSGD", "YVIG",
                "RTSY", "SAQFLRNN", "GVIG", "TSYR", "GVKT", "GTV", "NYN",
                "QFRDSVVIY", "GFDL", "DMN", "LIDNLLVR", "LGVGFGMDGKG",
                "YVSNITVED", "AGSGAYLLTHE", "VFTNIAIID", "QIYI", "RVNGLRL",
                "TIDAPNSTVSGITG", "INVANL", "NIRANS", "GYDSAAIKL", "KTL",
                "SGALYSHI", "AYTQLTAIS", "TPDAVSLKVN", "GAE", "VPD",
                "DSSCFLPYWE", "SLKALVK", "LVRLTLA"]
```

Via de geavanceerde datastructuur `Counter` kunnen we makkelijk de frequenties van de verschillende aminozuren bepalen.

```
from collections import Counter

freq_az_P22 = Counter(P22eiwit)

for AZ, freq in freq_az_P22.items():
    print(AZ, " : ", freq)
```

Ditto voor de β-platen:

```
freq_az_betaplaten = Counter()  # initieer een counter object

for betaplaat in beta_platen:
    # voeg de individuele platen toe
    freq_az_betaplaten.update(betaplaat)

# kijk voor aminozuur P (proline)
print("P :", freq_az_betaplaten["P"])
```

> **Optionele programmmeeropdracht**: bereken met de computer de kansverhoudingen zoals je op papier gedaan hebt. Bereken de kans dat peptide „YSIEADKK" een β-plaat is volgens de Naive bayes methode.

En verder...

De concepten die je in deze praktische sessie geleerd hebt zijn eenvoudig en kunnen zeer nuttig zijn in de praktijk. Deze methode is een vereenvoudigede versie van de **Chou-Fasman** methode om secundaire structuren te voorspellen. Er bestaan echter ook veel complexere methoden om eiwitten te bestuderen. Misschien vind je onze methode van het glijdend venster nogal onelegant. Een veel krachtigere methode om secundaire structuren te bepalen is via *verborgen Markovketens* (Engels: Hidden Markov Chains) die op een slimme manier eiwit- en DNA-sequenties kunnen labellen.

Daarenboven staat onderzoek in de bio-informatica nooit stil en zijn er zelfs grote bedrijven in geïnteresseerd, net omdat computers ons veel kunnen bijleren over biologie. Een zeer recent voorbeeld is Deepmind, een bedrijf dat onder Google werkt. Recent werk van hen gebruikt complexe artificiële intelligentie om de tertiaire structuur van een eiwit accuraat te voorspellen. Hun ontwikkelde methode *AlphaFold* is de eerste in zijn soort, maar zal waarschijnlijk niet de laatste zijn. Net zoals we in dit project gedaan hebben, werd hierbij een model gefit aan een databank met geannoteerde voorbeelden. Wij hebben echter met een model gewerkt met een twintigtal parameters, in de praktijk zijn het er miljoenen of miljarden.

Verspreiding van ziektes doorheen sociale netwerken

In dit project zullen we bestuderen hoe ziektes zich kunnen verspreiden via een (sociaal) netwerk. We zullen onderzoeken hoe de structuur van een netwerk een invloed kan hebben op hoe snel een ziekte doorgegeven wordt. Ten slotte zullen we het effect van vaccinatie bekijken.

Besmettelijke ziektes

In de geschiedenis van de mensheid[11] heeft niets zoveel mensen gedood als besmettelijke ziekten. De ziekte die de meeste mensen heeft gedood is waarschijnlijk tuberculose, waarbij 1000 miljoen mensen in alleen de 19de en 20ste eeuw werden gedood. De ziekte die het snelst gedood heeft, is de „Spaanse griep" epidemie die 50 tot 100 miljoen mensen heeft gedood tussen 1918 en 1920. De ziekte met het grootste evenredige dodental blijkt de Zwarte Dood, die 20% van de wereldbevolking en zelfs 50% van de Europese bevolking in de 14e eeuw doodde. Nog erger dan die ramp was het dodental op het Amerikaanse continent na de Europese kolonisatie. Er wordt geschat dat 90% van de inheemse bevolking geëlimineerd werd door de door Europeanen meegebrachte ziektes. De inheemse bevolking was namelijk nog nooit eerder met deze ziektes in contact gekomen en had dus nog niet de gelegenheid gehad om resistentie te ontwikkelen tegen deze ziektes.

Tegenwoordig veroorzaken besmettelijke ziektes minder doden door betere medische kennis, technieken en middelen. Echter blijven ze een ernstig probleem voor de volksgezondheid. Toch zijn 3 van de 10 belangrijkste doodsoorzaken wereldwijd besmettelijke ziekten. Naast de verbeteringen van de moderne wereld zijn er ook nieuwe uitdagingen voor het stoppen van epidemieën. Nu kan een besmette persoon op een vliegtuig stappen en in een paar uur tijd een ziekte naar een ander continent verspreiden.

[11] In deze nota's leggen we de nadruk op de verspreiding van humane ziektes. Dezelfde modellen worden echter ook gebruikt om ziektes tussen dieren (zoals verspreiding van *Myxomatosis* bij konijnen) of zelfs tussen planten (bijvoorbeeld verspreiding van *Phytophthora infestans*, een aardappelziekte die verantwoordelijk was voor de Ierse hongersnood).

Figuur 8: Verschillende vliegtuigroutes tussen luchthavens. De wereld is sterk geconnecteerd en ziektes kunnen zich nu veel sneller verspreiden.

Ziekteverspreidingsmodellen

Bij de uitbraak van een besmettelijke ziekte is het belangrijk om inzicht te krijgen in hoe deze ziekte zich de volgende dagen en weken kan verspreiden. Dit zal helpen om het optimale gebruik van middelen en mensen te plannen om de ziekteverspreiding een halt toe te roepen. Bovendien, voordat we deze maatregelen ten uitvoer leggen, moeten we weten hoe ze het verloop van de ziekte kunnen beïnvloeden, zodat de meest efficiënte en goedkoopste maatregelen eerst geïmplementeerd kunnen worden.

Om deze en andere belangrijke vragen te beantwoorden, worden *ziekteverspreidingsmodellen* vaak gebruikt. Deze modellen zijn gebaseerd op wiskundige vergelijkingen die beschrijven hoe een besmettelijke ziekte zich doorheen de tijd en/of ruimte verspreidt. Ze kunnen worden gebruikt om relevante vragen te beantwoorden, zoals:

- Zonder tussenkomst, zal de ziekte uitsterven of ongecontroleerd verspreiden?
- Hoeveel mensen moeten gevaccineerd worden om de verspreiding te stoppen?
- Hoe beïnvloedt het gedrag van mensen of dieren de verspreiding van ziektes?
- Welk effect zullen verschillende quarantainemaatregelen hebben?

Aangezien het verloop doorheen de tijd cruciaal is in het geval van ziekteverspreiding, zijn bijna alle

ziektespreidingsmodellen *dynamisch*[12] van aard. Maar sinds de ontwikkeling van de eerste ziektespreidingsmodellen in de eerste helft van de 20e eeuw, zijn bijna alle mogelijke modeltypes gebruikt om de dynamiek van besmettelijke ziektes te beschrijven.

In dit project zullen we twee van de belangrijkste en meest gebruikte types van ziektespreidingsmodellen bestuderen.

Het SIR-model

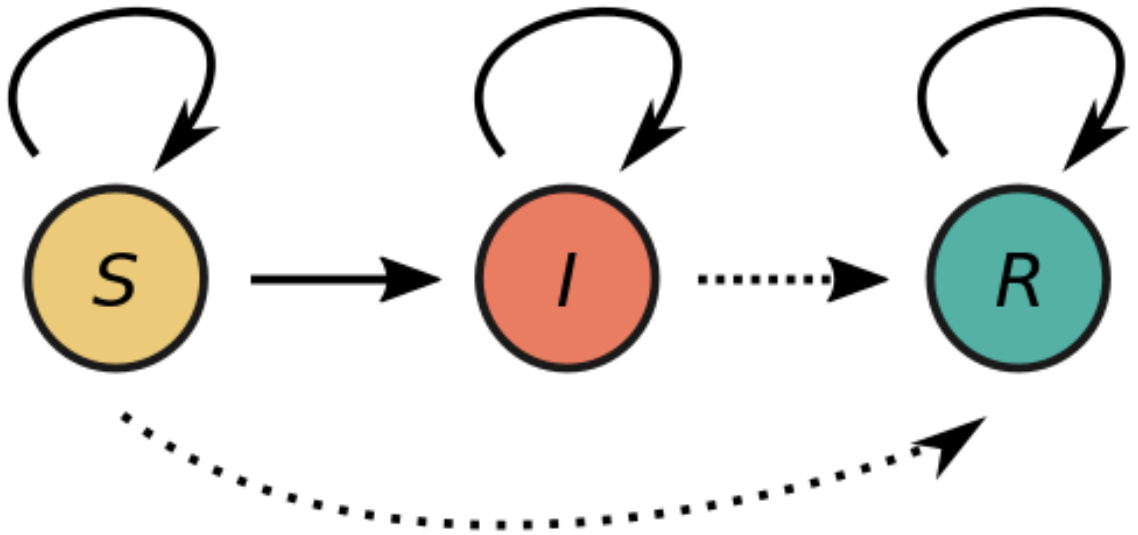
Een van de eenvoudigste manieren om ziekteverspreiding in een gemeenschap te modelleren is aan de hand van het SIR-model. SIR staat voor *Susceptible* (vatbaar), *Infected* (geïnfecteerd) en *Resistant* (resistent of hersteld), de drie types individuen die in een gemeenschap voorkomen. Meestal zijn de individuen gewoon mensen, maar dit model kan ook aangewend worden om ziekteuitbraak bij knaagdieren, vogels of zelfs planten te modelleren. Het SIR-model bestaat uit drie vergelijkingen die de veranderingen van het aantal individuen in een bepaalde groep beschrijven. De variabelen die de toestand beschrijven zijn:

- $S(t)$: het (relatief) aantal vatbare individuen op tijdstip t;
- $I(t)$: het (relatief) aantal geinfecteerde individuen op tijdstip t;
- $R(t)$: het (relatief) aantal resistente individuen op tijdstip t.

In deze beschrijving maken we een eerste grote vereenvoudiging van de werkelijkheid. We nemen aan dat elk van deze variablen reëelwaardig zijn en dat het aantal individuen in elke groep continu kan variëren. In werkelijkheid is het aantal geïnfecteerden of vatbare individuen een natuurlijk getal, je bent immers besmet of je bent het niet. Modelleerders werken echter graag met continue variablen omdat ze dan de technieken van wiskundige analyse kunnen gebruiken.

> **Vraag 1**: Onder welke omstandigheden gaat deze continue benadering ongeveer op? Wanneer niet?

[12] Een dynamisch model modelleert een verandering in de tijd.



Figuur 9: Visule voorstelling van het SIR-model. Een vatbaar individu (toestand S) kan geïnfecteerd worden (toestand I), weergegeven door de volle pijlen. Een geïnfecteerd individu kan immuun worden en vatbare individuen kunnen geïmmuniseerd worden, weergegeven door de pijlen met stippellijnen. In dit project zullen we deze overgangen niet beschouwen.

Deze drie variabelen worden aan elkaar gelinkt aan de hand van drie differentiaalvergelijkingen (die elk een verandering in de tijd beschrijven). Hierin nemen we aan dat de grootte van de populatie ongewijzigd blijft. We nemen dus aan dat, gedurende de tijdspanne die het model beschrijft, er niemand geboren wordt en ook niemand sterft. We zullen ons hier dus beperken tot de verspreiding van een relatief onschuldige ziekte zoals een verkoudheid. De drie vergelijkingen zijn als volgt:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta\, S(t)\, I(t)$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta\, S(t)\, I(t) - \gamma\, I(t)$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma\, I(t)$$

Elke vergelijking vertelt ons hoe het aantal mensen in elke groep wijzigt doorheen de tijd. Daaruit kunnen we ook berekenen hoeveel mensen zich op een bepaald moment bevinden in elke groep. De vergelijkingen zijn gekoppeld via de *overgangssnelheden*. Elke overgangssnelheid vertelt ons hoe waarschijnlijk het is om van de ene naar de andere groep over te gaan.

De overgangssnelheid van vatbaar (S) naar geïnfecteerd (I) hangt af van het contact tussen een vatbare persoon en een geïnfecteerd persoon. We noemen deze contactsnelheid β. De kans dat de ziekte overgedragen wordt tijdens een contact tussen een vatbare en een geïnfecteerde persoon is dus βI. Het aantal vatbare personen (S) vermindert dus met deze snelheid op elk tijdstip.

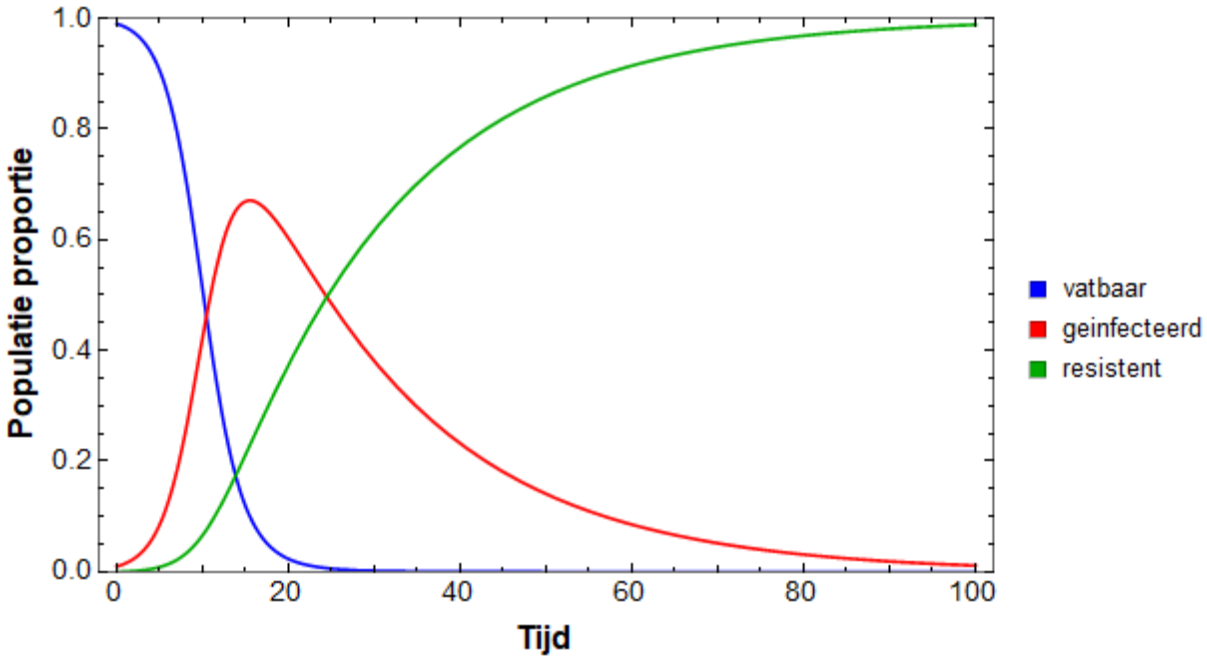
De overgangssnelheid van geïnfecteerd (I) naar resistent (R) hangt alleen af van de snelheid van herstel, die we γ noemen. Het aantal geïnfecteerde personen vermindert dus met deze snelheid op elk tijdstip.

> **Vraag 2**: Kan je aantonen dat het totaal aantal individuen in de populatie $(S(t) + I(t) + R(t))$ constant zal blijven?

Bij het vaststellen van deze overgangssnelheiden hebben we een andere belangrijke vereenvoudiging gemaakt. We nemen aan dat elke persoon in de populatie een gelijke waarschijnlijkheid heeft om in contact te komen met elke andere persoon. Anders gezegd, we nemen aan dat de populatie perfect gemengd is. In sommige gevallen kan deze vereenvoudiging passen bij de realiteit, bijvoorbeeld als we willen bijhouden hoe de griep zich door een fuif verspreidt.

Het SIR-model kan niet exact worden opgelost, zoals veel differentiaalvergelijkingen die optreden in de biologische wetenschappen. We moeten dus een *numerieke benadering* van de oplossing vinden. Dit betekent dat we een algoritme zullen gebruiken om een geschatte maar nauwkeurige oplossing te vinden. Er zijn verschillende mogelijkheiden om dit te doen. We zouden ons continue probleem kunnen vervangen door een discrete tegenhanger. Dit zou ons toelaten bepaalde numerieke methoden te gebruiken om een benaderende oplossing te krijgen. Anderzijds kunnen we een iteratieve methode gebruiken. Uitgaande van een initiële schatting, maken iteratieve methoden opeenvolgende benaderingen die stapsgewijs convergeren naar de exacte oplossing.

Met behulp van computers is het gemakkelijk om op deze manier een numerieke oplossing voor het SIR-model te vinden. Vanuit deze oplossing kunnen we leren hoe de verschillende variabelen in de loop van de tijd veranderen. Om dit te doen, vertrekken we van een beginvoorwaarde: het is logisch om te beginnen met een populatie met nul resistente personen, een paar geïnfecteerde personen en de rest vatbaar. Vervolgens kunnen we onze numerieke oplossing gebruiken om het aantal mensen in elke groep op elke tijdstap te berekenen. Als we een plot maken, zullen we de dynamiek zien die in de Figuur 10 wordt getoond.



Figuur 10: Een simulatie van een oplossing van het stelsel differentiaalvergelijkingen die het standaard SIR-model voorstellen.

> **Vraag 3**: Een epidemie wordt **uitbreidend** genoemd als het aantal geïnfecteerden toeneemt. Wanneer is de epidemie uitbreidend? Op het moment wanneer I verandert van toenemend naar afnemend, wat kun je zeggen over de verandering van I? (**hint**: kijk naar de vorm van $\frac{dI}{dt}$)

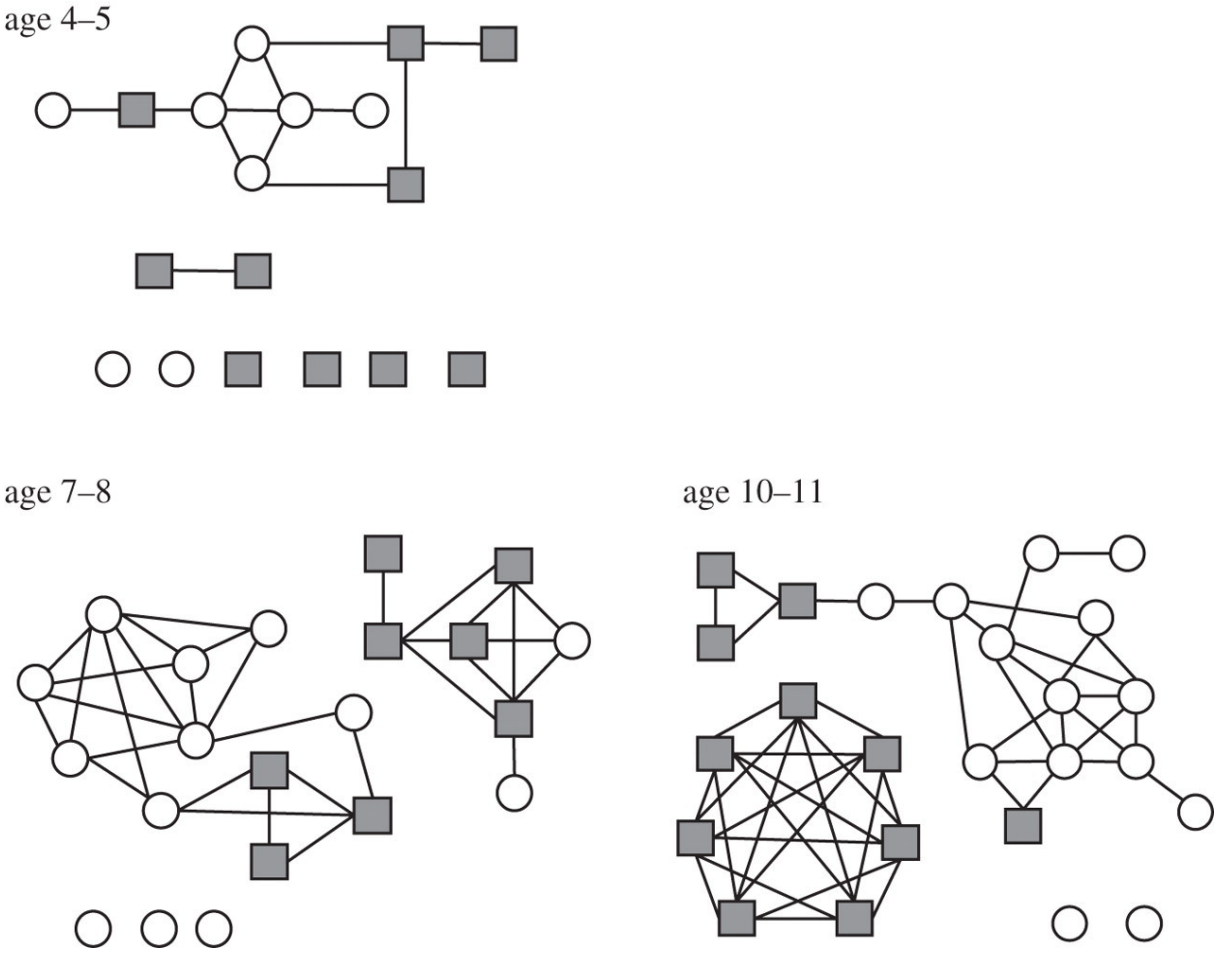
Sociale netwerken

Het standaard SIR-model maakt de onrealistische veronderstelling dat twee willekeurige individuen telkens dezelfde kans hebben om met elkaar in contact te komen en zo mogelijks een ziekte door te geven [13]. In werkelijkheid gaat natuurlijk niet iedereen met dezelfde mensen om. We hebben allemaal mensen waar we meer mee omgaan (meer in contact mee komen) dan met anderen. Het geheel van wie met wie in contact staat wordt een *sociaal netwerk* genoemd (denk aan Facebook). Het lijkt evident dat de structuur van zo'n netwerk een sterke invloed zal hebben op de dynamiek van de ziekteverspreiding. In deze sectie zullen we bekijken hoe we een netwerk wiskundig kunnen beschrijven.

[13] Welk type contact hier ook voor nodig zou zijn.

Een voorbeeld

Hieronder zie je een voorbeeld van enkele netwerken. De punten vertegenwoordigen de leerlingen en worden *knopen* genoemd. De contacten tussen leerlingen worden weergegeven door lijnsegmenten tussen knopen, en worden *bogen* genoemd. We zeggen dat twee knopen met elkaar *verbonden* zijn als ze met een boog geconnecteerd worden. Hier gaan we ervan uit dat een knoop niet verbonden kan zijn met zichzelf[14]. Ook is er maar maximaal één boog mogelijk tussen twee knopen. De *graad* van een knoop is het aantal bogen dat ermee verbonden zijn.



Figuur 11: Voorbeelden van gekleurde grafen die netwerken tussen kinderen van verschillende leeftijden voorstellen (bron).

Zoals je ziet wordt een netwerk of een graaf vaak voorgesteld in een figuur waar cirkels (of andere elementen) de knopen voorstellen die geconnecteerd zijn door lijnen, de bogen. Deze figuren zijn niet

[14]We gaan ervan uit dat je niet kan bevriend zijn met jezelf.

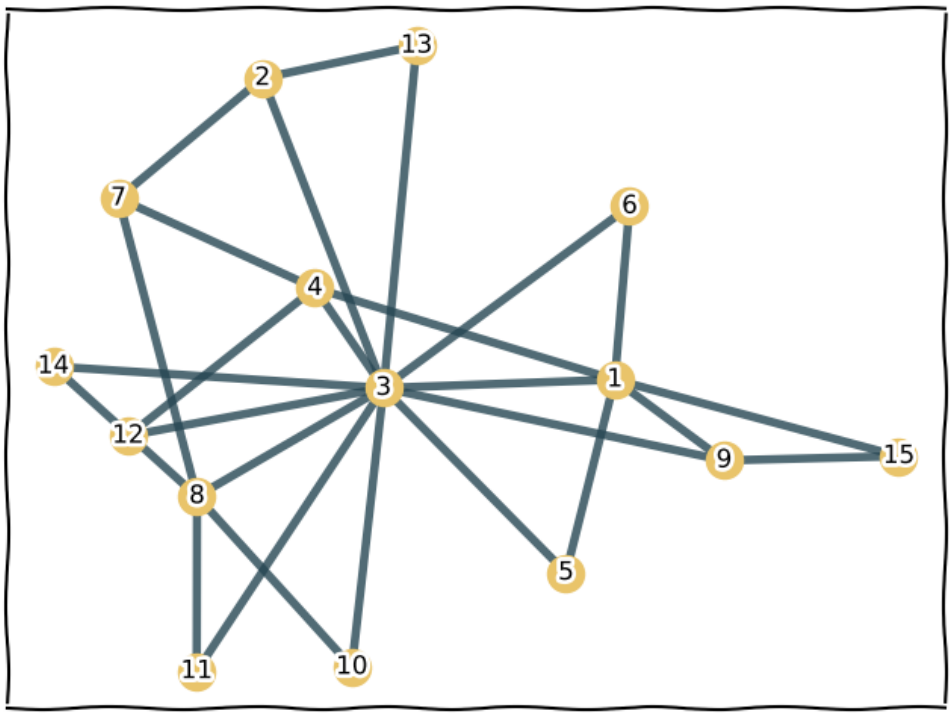
uniek: eenzelfde netwerk kan vaak op verschillende manieren voorgesteld worden. Soms hebben de knopen ook een kleur, bijvoorbeeld om geslacht te duiden in een sociaal netwerk. In dat geval spreekt men van een *gekleurde graaf*.

> **Vraag 4**: Beschrijf het verschil tussen de sociale netwerken tussen kinderen van verschillende leeftijden.

Een figuur is nuttig om te bekijken hoe het netwerk eruitziet. Om er berekeningen mee te doen zijn er echter andere voorstellingen nodig. Een graaf kan wiskundig voorgesteld worden in een matrix die een *bogenmatrix* (Engels: adjacency matrix) genoemd wordt. Als het aantal knopen in de graaf n is, dan is de bogenmatrix een vierkante matrix met dimensies $n \times n$. Het element $A_{ij} = 1$ als de knopen i en j verbonden zijn, en $A_{ij} = 0$ als ze niet verbonden zijn[15]. De bogenmatrix linkt graaftheorie met matrixtheorie!

Het sociaal netwerk dat we beschouwen wordt weergegeven in onderstaande figuur. De knopen, hier personen, zijn genummerd om ze makkelijk te kunnen identificeren. We houden geen rekening met geslacht of andere attributen. We zullen hier een ziekte-uitbraak op simuleren!

[15] In het echte leven hebben de meeste mensen in een populatie geen contact met elkaar (denk aan het sociaal netwerk van een hele stad). De graaf is dus verre van *volledig verbonden* (elk paar knopen is verbonden) en de elementen van de bogenmatrix bestaat grotendeels uit nullen. In deze gevallen kan het soms beter zijn om een *verbindingslijst* te gebruiken. Dit is een lijst met dimensies $m \times 2$ waarbij m het aantal bogen is, en elke rij bevat een koppel knopen die verbonden zijn. Afhankelijk van het specifieke netwerk dat we bestuderen en wat we ermee willen doen, kan de ene of de andere datastructuur efficiënter zijn.



Figuur 12: Een sociaal netwerk tussen vijftien personen.

> **Oefening 1**: voltooi de bogenmatrix voor het sociale netwerk.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
5	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
6	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
7	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
8	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
9	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
10	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
11	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
12	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
13	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
14	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
15	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Gradenverdeling

> **Vraag 5**: Wie zal er eerder een verkoudheid oplopen: persoon 3 of 15 ?

Een graaf is een complexe wiskundige structuur. Een gegeven knoop in een graaf wordt gekarakteriseerd door zijn bogen en dus ook graad. Echter, belangrijke eigenschappen van de graaf zijn *emergent*, dit wil zeggen dat ze enkel te verklaren zijn door de graaf in zijn geheel en niet enkel de individuele onderdelen.

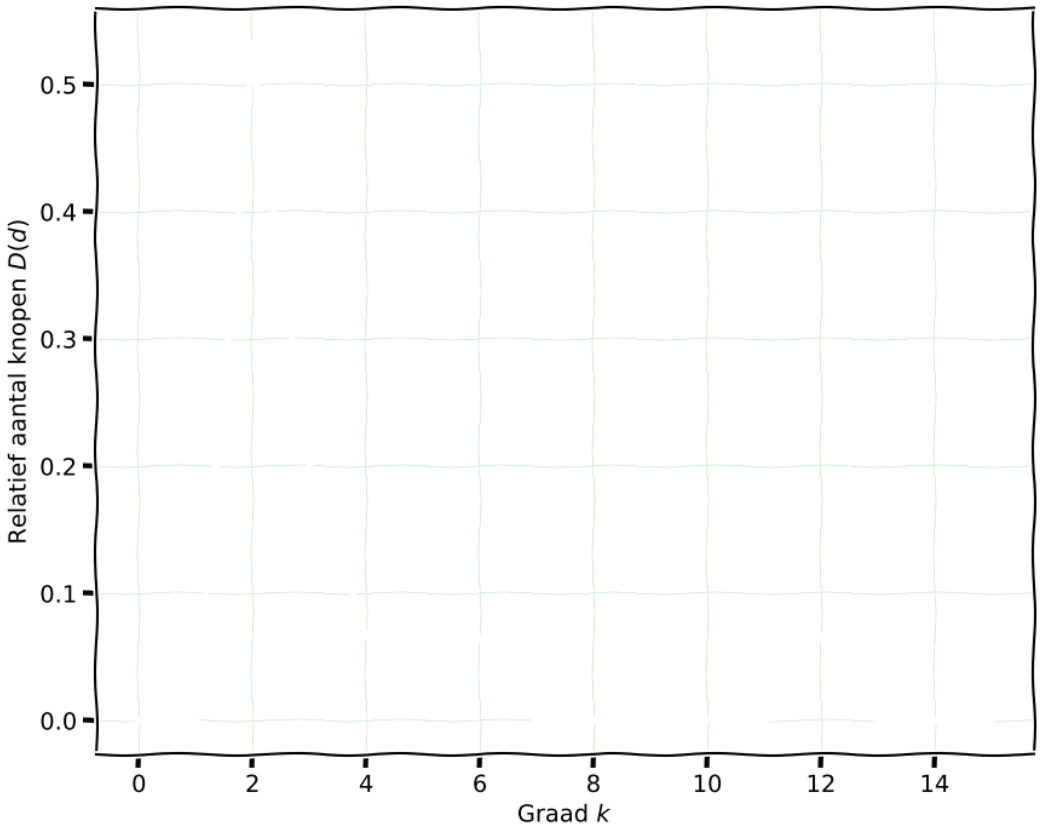
Als we iets willen leren over een netwerk, welke informatie kunnen we dan bekijken? Het zou zeer interessant zijn om te weten hoe verbonden het netwerk is. Hoe kunnen we bepalen in welke mate de knopen met elkaar verbonden zijn? We zouden naar de gemiddelde knoopgraad kunnen kijken, maar dit zou ons niet veel zeggen. In plaats daarvan kunnen we de fractie van de knopen met graad k tellen. Deze plot wordt de *gradenverdeling* (Engels: *degree distribution*) genoemd en kan ons veel vertellen

over de structuur van een netwerk. Wiskundig drukken we de gradenverdeling uit als

$$D(k) = \text{fractie van de knopen met } k \text{ bogen (graad } k\text{)}\,.$$

> **Oefening 2**: Bereken en plot de gradenverdeling van het sociaal netwerk. Vul eerst de aantallen (niet fracties) in onderstaande tabel in, normalizeer deze aantallen en teken dan de plot (met genormalizeerde waarden).

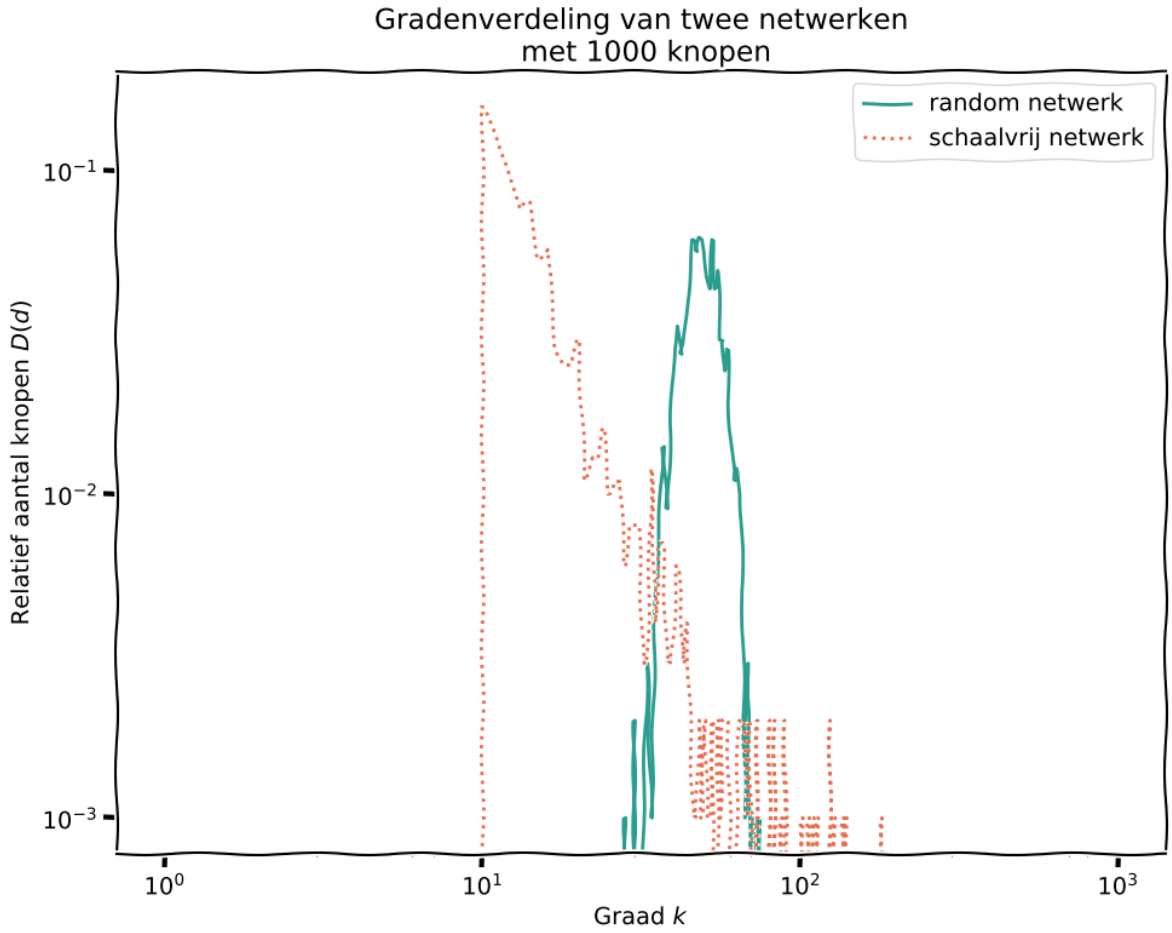
k	1	2	3	4	5	6	7	8	9	10
aantal knopen met graad k	…	…	…	…	…	…	…	…	…	…



Figuur 13: Schets hier de gradenverdeling. Vergeet niet te normalizeren!

Twee typevoorbeelden van netwerken

Er zijn vele verschillende types van netwerken. We beschouwen twee belangrijke: **willekeurige** netwerken en **schaalvrije** netwerken.



Figuur 14: Illustraties van gradenverdelingen van verschillende types van netwerken.

Willekeurige netwerken

Een *willekeurig netwerk* heeft eigenschappen die willekeurig worden bepaald, zoals het aantal knopen, het aantal bogen en de verbindingen tussen knopen en bogen. We kunnen een dergelijk netwerk beschrijven als $G(n, p)$ waarbij n het aantal knopen is en p, een getal tussen 0 en 1, de kans dat er een boog tussen een willekeurig paar knopen is.

Bij een willekeurig netwerk is het verwacht aantal graden per knoop gelijk aan

$$n \cdot p \,.$$

De meeste knopen hebben een graad dicht bij dit gemiddelde. **In een (groot) willekeurig netwerk vind je zelden een knoop met extreem veel of extreem weinig bogen**[16].

> **Vraag 6** Binnen een random netwerk heeft elke knoop ongeveer dezelfde graad (iedereen heeft ongeveer evenveel vrienden in een sociaal netwerk). Denk je dat dit een realistische assumptie is voor veel netwerken?

Schaalvrije netwerken

Een netwerk wordt *schaalvrij* (Engels: scale-free) genoemd als de gradenverdeling ongeveer aan volgdende vorm voldoet, een zogenaamde *power law*[17]

$$D(k) \propto \frac{1}{k^a} \,,$$

met a een exponent die typisch kan verschillen van netwerk tot netwerk. **In een schaalvrij netwerk hebben enkele knopen een hoge graad en zijn er veel knopen met een lage graad.** Schaalvrije netwerken onstaan door een aggregatieproces waarbij *de rijken rijker worden*: wanneer nieuwe knopen aan een netwerk toegevoegd worden, gaan deze bij voorkeur verbindingen aan met knopen met een reeds hoge graad.

Schaalvrije netwerken komen overal voor:

- netwerken van filmsterren (knopen) die samen in een film gespeeld hebben (bogen);
- netwerken van het internet: links (bogen) tussen websites (knopen);
- netwerken die interacties (bogen) weergeven tussen eiwitten (knopen).

> **Vraag 7**: Stel je een sociaal netwerk voor van drie vrienden waar gradueel nieuwe mensen aan geïntroduceerd worden. Kan je je een scenario voorstellen waarbij een schaalvrij netwerk bekomen zou worden?

Verspreiding van een ziekte doorheen een netwerk

Laat ons nu kijken hoe we het SIR-ziekteverspreidingsmodel kunnen vertalen naar de taal van netwerken. Aan de hand van een algemeen netwerk zullen we een veel realistischer model opstellen. Geen

[16] Om precies te zijn, de kans dat een knoop in een netwerk met n knopen exact m bogen heeft wordt gegeven door $p^m \, (1-p)^{\binom{n}{2}-m}$. Dit volgt uit de binomiale verdeling. Hier is $\binom{n}{2}$ de binomiaalcoëfficient $\binom{n}{2} = \frac{n(n-1)}{2}$. Dit is het aantal mogelijke manieren waarop je twee knopen kunt kiezen uit n.

[17] $\propto$ wil zeggen „evenredig aan”.

continue benadering meer! Vreemd genoeg sluit dit model niet enkel dichter aan bij de werkelijkheid, maar is het ook veel eenvoudiger om te bevatten en te simuleren. We kunnen een exacte oplossing bekomen zonder zelfs maar afgeleiden of andere geavanceerde wiskundige technieken nodig te hebben!

Ziektedynamiek op een netwerk

In plaats van het aantal S, I en R individuen doorheen de tijd bij te houden zoals bij het standaard SIR-model, zullen we voor elke knoop in het netwerk zijn of haar toestand bijhouden. Ook de tijd zal niet meer continu variëren maar zal nu in discrete stappen voorbij gaan: $t = 0, 1, 2, 3, ...$ De toestanden van het model worden dus beschreven door $N_i^t \in \{S, I, R\}$. Dit wil zeggen dat knoop i op tijdstip t de toestand S (vatbaar), I (geïnfecteerd) of R (resistent) kan hebben. De verandering in toestanden voor de knopen beschrijven we aan de hand van enkele eenvoudige regels.

Vatbare en geïnfecteerde mensen

Laten we ons eerst beperken tot vatbare en geïnfecteerde individuen. We gaan ervan uit dat vatbare individuen geïnfecteerd kunnen worden, maar geïnfecteerde individuen niet meer kunnen genezen. Beschouw volgende regels:

1. Indien een knoop op tijdstip t in toestand S zit en al zijn buren eveneens in toestand S zitten, blijft de knoop op tijdstip $t + 1$ in toestand S.
2. Indien een knoop op tijdstip t in toestand S zit en minstend één van zijn buren eveneens in toestand I zit, verandert de knoop op tijdstip $t + 1$ naar toestand I.
3. Indien een knoop op tijdstip t in toestand I zit blijft de knoop op tijdstip $t + 1$ in toestand I.

> **Oefening 3** Gebruik het sociale netwerk dat je hebt gekregen om de verspreiding van een ziekte te modelleren.
> 1. Elk individu begint als vatbaar. Kies een persoon om de eerste geïnfecteerde te worden en kleur deze in.
> 2. Op elke tijdstip, ga één voor één door de buren van een geïnfecteerde persoon en laat hen ook geïnfecteerde raken volgens de bovenstaande regels. Vul de tabel in om de spreiding over de tijd te volgen.
> 3. Herhaal totdat het netwerk niet meer verandert.
> 4. Volgens de tabel, plot het aantal geïnfecteerden op elke tijdstip.

knoop	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
1	…	…	…	…	…	…	…	…	…
2	…	…	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…	…	…
5	…	…	…	…	…	…	…	…	…
6	…	…	…	…	…	…	…	…	…
7	…	…	…	…	…	…	…	…	…
8	…	…	…	…	…	…	…	…	…
9	…	…	…	…	…	…	…	…	…
10	…	…	…	…	…	…	…	…	…
11	…	…	…	…	…	…	…	…	…
12	…	…	…	…	…	…	…	…	…
13	…	…	…	…	…	…	…	…	…
14	…	…	…	…	…	…	…	…	…
15	…	…	…	…	…	…	…	…	…
totaal aantal geinfecteerden	…	…	…	…	…	…	…	…	…

> **Oefening 4**: Herhaal de laatste oefening met een ander startpunt: kies een persoon met minder of meer contacten. Vul de tabel in en plot het aantal geïnfecteerden op elke tijdstip. Hoe verandert de ziekteverspreiding ?

knoop	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
1	…	…	…	…	…	…	…	…	…
2	…	…	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…	…	…
5	…	…	…	…	…	…	…	…	…
6	…	…	…	…	…	…	…	…	…

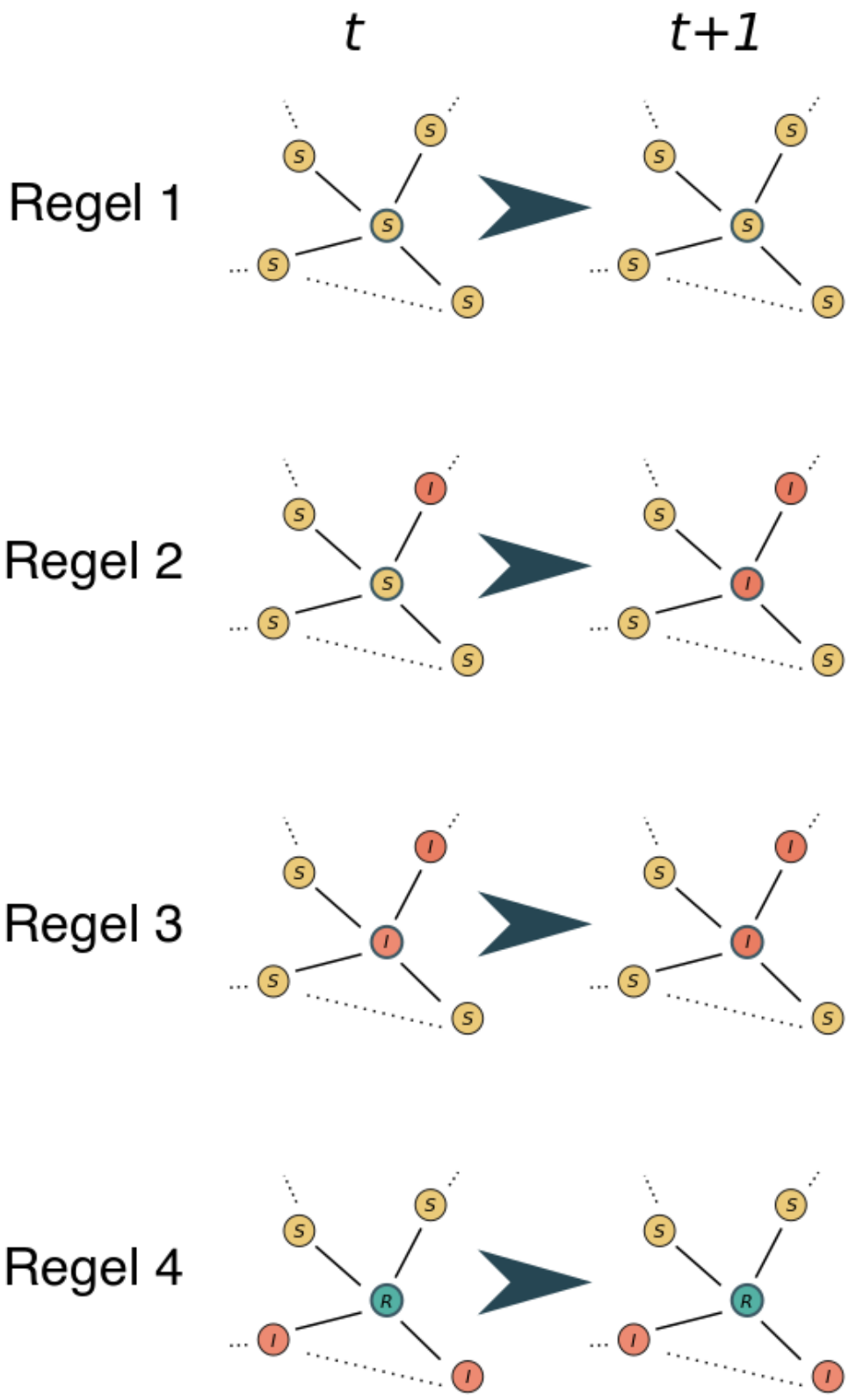
knoop	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
7	…	…	…	…	…	…	…	…	…
8	…	…	…	…	…	…	…	…	…
9	…	…	…	…	…	…	…	…	…
10	…	…	…	…	…	…	…	…	…
11	…	…	…	…	…	…	…	…	…
12	…	…	…	…	…	…	…	…	…
13	…	…	…	…	…	…	…	…	…
14	…	…	…	…	…	…	…	…	…
15	…	…	…	…	…	…	…	…	…
totaal aantal geinfecteerden	…	…	…	…	…	…	…	…	…

Immuniteit en vaccinatie

Het bovenstaande voorbeeld geeft ons een idee hoe we ziekteverspreiding op een netwerk kunnen modelleren, maar nu kunnen we enkele vereenvoudigingen verwijderen. Laten we nu *immuniteit* in onze populatie toestaan. Dit betekent dat sommige mensen niet geïnfecteerd kunnen worden. Hun immuniteit kan natuurlijk zijn (door een herstelling van een eerdere infectie) of kunstmatig (door een vaccin te krijgen). Nu kunnen individuen in het netwerk zich ook in toestand R (resistent) bevinden.

Beschouw nu de volgende regels (visueel voorgesteld in Figuur 15):

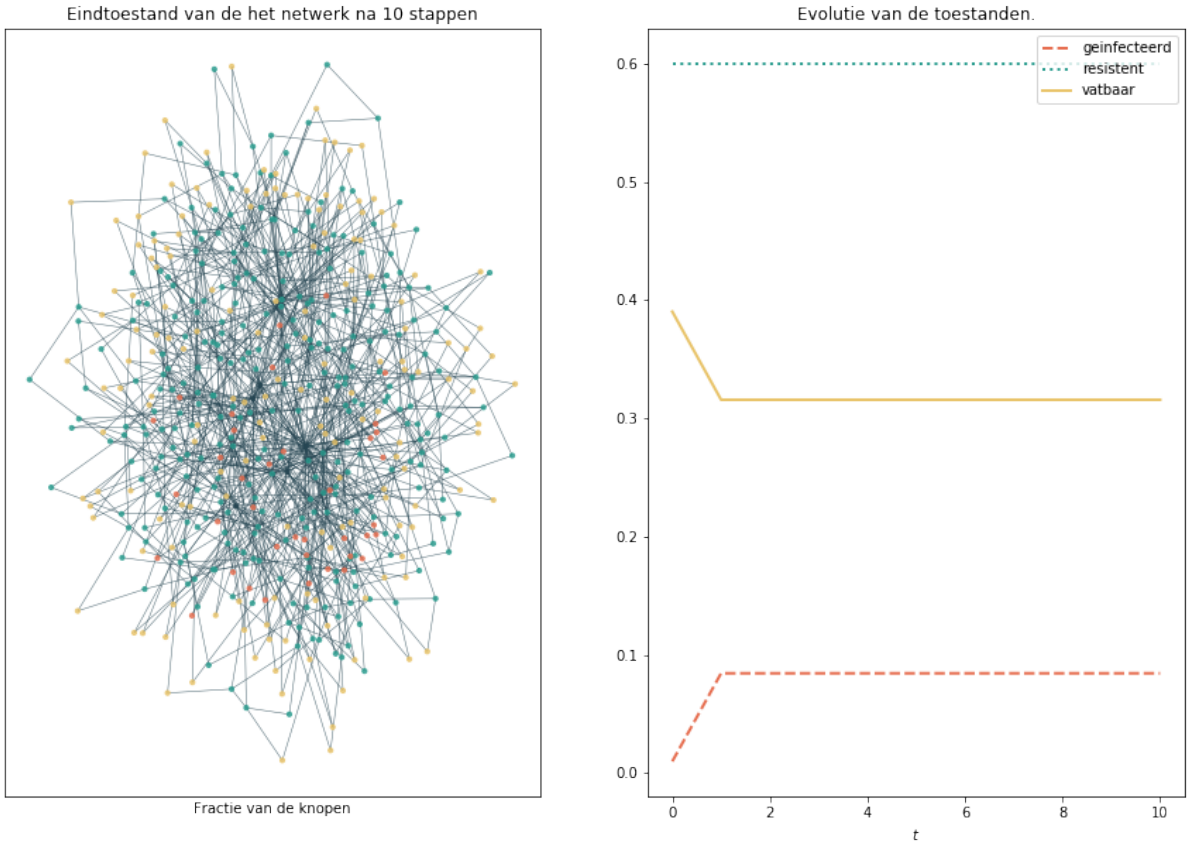
1. Indien een knoop op tijdstip t in toestand S of R zit en al zijn buren eveneens in toestand S of R zitten, verandert de knoop zijn toestand niet op tijdstip $t+1$.
2. Indien een knoop op tijdstip t in toestand S zit en minstens één van zijn buren eveneens in toestand I zit, verandert de knoop op tijdstip $t+1$ naar toestand I.
3. Indien een knoop op tijdstip t in toestand I zit blijft de knoop op tijdstip $t+1$ in toestand I.
4. Een knoop in toestand R blijft altijd in toestand R.



Figuur 15: Overzicht van de regels voor het SIR-model op een netwerk.

Sommige mensen kunnen door verschillende redenen niet immuun worden. Vaccins kunnen bijvoorbeeld niet gegeven worden aan jonge baby's of mensen met ernstige medische aandoeningen. In deze groep is *kudde-immuniteit* een belangrijke beschermingsmethode.

Kudde immuniteit betekent een indirecte bescherming tegen besmettelijke ziekten. Deze komt voor wanneer een groot percentage van de populatie immuun is tegen een infectie (door natuurlijke immuniteit of vaccinatie) en daardoor beschermen ze mensen die niet immuun zijn. Dit gebeurt omdat het grote aantal immune mensen de ziekteverspreiding vertraagt of zelfs stopt, want de verbindingen tussen zieke en vatbare mensen zijn geblokkeerd.



Figuur 16: Illustratie van kudde-immuniteit. Als voldoende mensen geïmmuniseerd zijn breidt de ziekte zich niet verder uit bij vatbare mensen. (links) Netwerk na 100 stappen. (rechts) Verdelingen van de toestanden in de tijd.

Als een bepaalde drempwelwaarde bereikt kan worden, zal de kudde-immuniteit een ziekte uit een populatie elimineren. Als deze eliminatie over de hele wereld bereikt wordt, kan het aantal infecties permanent tot nul teruggebracht worden. Dan kunnen we spreken van de *uitroeiing* van de ziekte. Het moet duidelijk zijn dat volledige uitroeiing zeer moeilijk te bereiken is. Veel ziekten zijn regionaal uitge-

roeid (bijvoorbeeld cholera in België), terwijl slechts twee ziekten wereldwijd uitgeroeid zijn: pokken en runderpest.

> **Computeroefening**: Laat ons overgaan naar simulaties op de computer. Je kan deze uitvoeren in de Jupyter notebooks, beschikbaar via de biowiskundedagen website. Via de interactieve widget kan je een netwerk van een bepaalde grootte genereren met 1 tot 10 geïnfecteerde personen (deze knopen zijn donkerblauw ingekleurd). Het netwerk dat verschijnt is na 10 tijdstappen. Ernaast zijn de fracties van de knopen in een bepaalde toestand geplot.
>
> - **zonder vaccinatie**: `frac_vac=0`
>
> 1. Hoeveel tijdstappen zijn er nodig voordat iedereen geïnfecteerd is?
> 2. Is er een verschil tussen hoe snel de ziekte zich verspreidt in een willekeurig of een shcaalvrij netwerk?
>
> - **met vaccinatie**: `frac_vac>0`, er is keuze tussen een bepaalde fractie individuen willekeurig te vaccinatie of de fractie meest geconnecteerde individuen te vaccineren.
>
> 1. Bekijk het effect van willekeurig vaccineren. Vanaf welke fractie worden ook vatbare individuen beschermd?
> 2. Wat is het verschil met gerichte vaccinatie?

Simuleren van een het SIR-model op een netwerk

Zoals eerder gezegd kunnen we via de verbindingsmatrix A een netwerk voorstellen. We kunnen een toestandsvector $\mathbf{x}$ gebruiken waar een 0 voorstelt dat die persoon vatbaar is en een 1 als die geïnfecteerd is.

```
from numpy import matrix
import numpy as np

# een matrix is een geneste lijst
# kan je de graaf tekenen hiervan?
A = matrix([[0, 1, 1, 0, 1],
            [1, 0, 1, 0, 1],
            [1, 1, 0, 0, 1],
            [0, 0, 0, 1, 0],
            [1, 0, 1, 0, 0]])

# toestandsvector met 1 persoon geïnfecteerd
x = matrix([[1, 0, 0, 0, 0]]).T
```

Simulatie kunnen we eenvoudig doen met een for-lus.

```
for t in range(5):  # 5 tijdstappen
    print("Tijdstip ",t, ": ", np.sum(x > 0), "geinfecteerden, x=",x.T > 0)
    x = A * x   # matrix-vector vermenidgvuldiging verspreidt de ziekte
```

> **Optionele programmmeeropdracht**: Kan je het model aanpassen zodat persoon 2 en 3 resistent zijn?

Ziektespreidingsmodellen in de praktijk

Epidemieën komen voortdurend voor en daarom gebruiken volksgezondheidsorganisaties over de hele wereld modellen om interventiestrategieën te ontwikkelen en te evalueren. Met behulp van simulaties kunnen ze snel de situatie beoordelen en belangrijke beslissingen nemen. Om een epidemie te herkennen en erop te reageren, hebben gezondheidswerkers informatie nodig die inherent onvoorspelbaar is (wat, waar, hoeveel gevallen, hoeveel zullen sterven, waar zal het zich verspreiden). De interacties die tot het uitbreken van een ziekte leiden zijn zeer complex, zodat de resultaten soms onverwacht of contra-intuïtief zijn. Er zijn modellen nodig om deze interacties te begrijpen en om de kwantitatieve voorspellingen te maken die volksgezondheidswerkers nodig hebben om te beslissen over interventiestrategieën.

Menselijk gedrag tijdens ziekte-uitbraken verandert vaak drastisch. Mensen vermijden drukke plaatsen of haasten zich naar drukke plaatsen zoals luchthavens of treinstations als ze proberen te ontsnappen aan de epidemie. Modellering kan gezondheidswerkers helpen dit soort effecten te voorzien en te begrijpen.

Modellen kunnen ook gebruikt worden om te bepalen hoe bestaansmiddelen toegewezen moeten worden om de beste kans te hebben om de verspreiding van de ziekte te stoppen - bijvoorbeeld, als vaccins beperkt zijn, welke groep mensen moet dan prioritair worden gevaccineerd? Wetenschappers kunnen modellen gebruiken om de uitkomsten van verschillende controlestrategieën te vergelijken. Modellen kunnen ook worden gekoppeld aan langetermijngegevens over het klimaat en klimaatvoorspellingen, om voorspellingen van uitbraken vele maanden in de toekomst te maken. Deze benadering wordt gebruikt om vaccinatiecampagnes te bepalen, bijvoorbeeld tegen influenza of mazelen.

Wetenschappers ontwikkelen hun begrip van ziekteverspreiding met behulp van gegevens zoals gedrags-, demografische en epidemische trends, maar het is vaak moeilijk om betrouwbare gegevens te verzamelen en voor veel ziekten missen we nog steeds belangrijke informatie over hoe ze zich verspreiden. Modellering kan ook in deze gevallen helpen, omdat wetenschappers verschillende hypotheses kunnen testen om te proberen de hiaten in hun kennis in te vullen.