Probabilistic representation of gene regulatory networks

Ziele des Vortrags

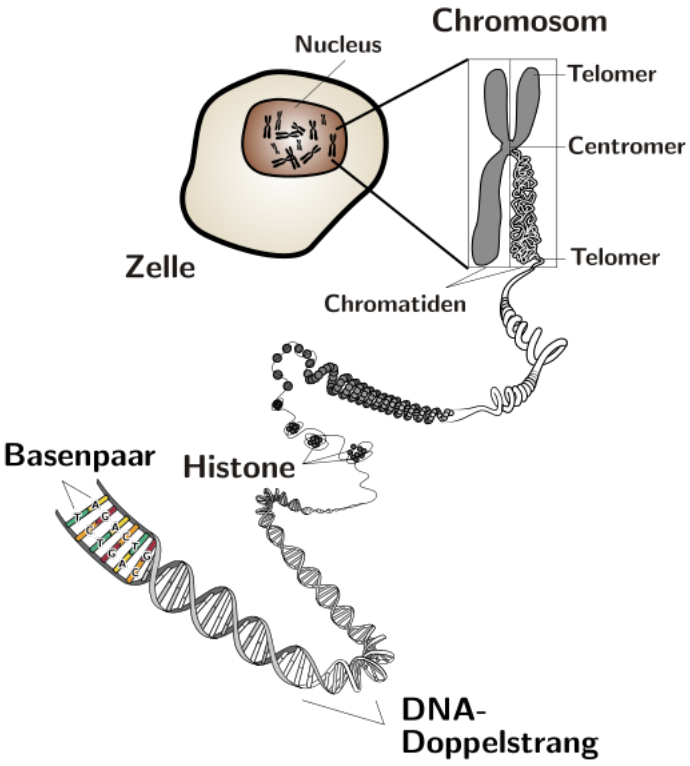
Worum soll es in diesem Vortrag gehen?

Ziel des Vortrags ist es, einen Einblick in die mathematische Modellierung eines **Gen Regulations Netzwerkes** (in Zukunft mit **GRN** abgekürzt) zu geben. Dafür erscheint es sinnvoll zunächst einige biologische Aspekte zu erläutern, um den Bezug zwischen Mathematik und der Realität zu verstehen. Natürlich wird lediglich ein Einblick in die biologische Funktionsweise der Netzwerke gegeben, da die genaue Betrachtung den Rahmen eines Vortrages übersteigen würde.

Biologische Grundlagen

Begrifflichkeiten

Zelle: Eine Zelle besteht aus vielen unterschiedlichen Bestandteilen. Der für unser Modell wichtige Teil ist der Zellkern. In diesem befindet sich in einem Stoffgemisch mit verschiedenen Proteinsäuren die DNA. Die verschiedenen Säuren (auch die DNA) im Zellkern beeinflussen sich gegenseitig und es kommt zu chemikalischen Reaktionen. Wie stark und wie oft solche Reaktionen stattfinden hängt unter anderem von der Konstellation des Gemisches und der Konzentration der Stoffe in dem Gemisch ab.



DNA: Die DNA(Desoxyribonukleinsäure) ist ein Molekül, bei dem sich entlang eines in schraubenförmiger Doppelhelix gewundenen Phosphatrückgrats jeweils 2 unterschiedliche Paare von Basen angelagert haben. Hierbei ist stets ein Guanin mit einem Cytosin und ein Adenin mit einem Thymin gepaart. Da die jeweiligen Partner immer gleich sind, reicht es diese durch eine Buchstabenkette entlang einer der beiden Helices zu beschreiben. So ergibt sich eine Buchstabenkette aus dem Alphabet $\alpha = G, A, C, T$. Nicht alle Teile der DNA sind wichtig. Es wird unterschieden zwischen Exons und Introns. Exons entsprechen den codierenden Sequenzen (Genen) und Introns den nicht codierenden Sequenzen

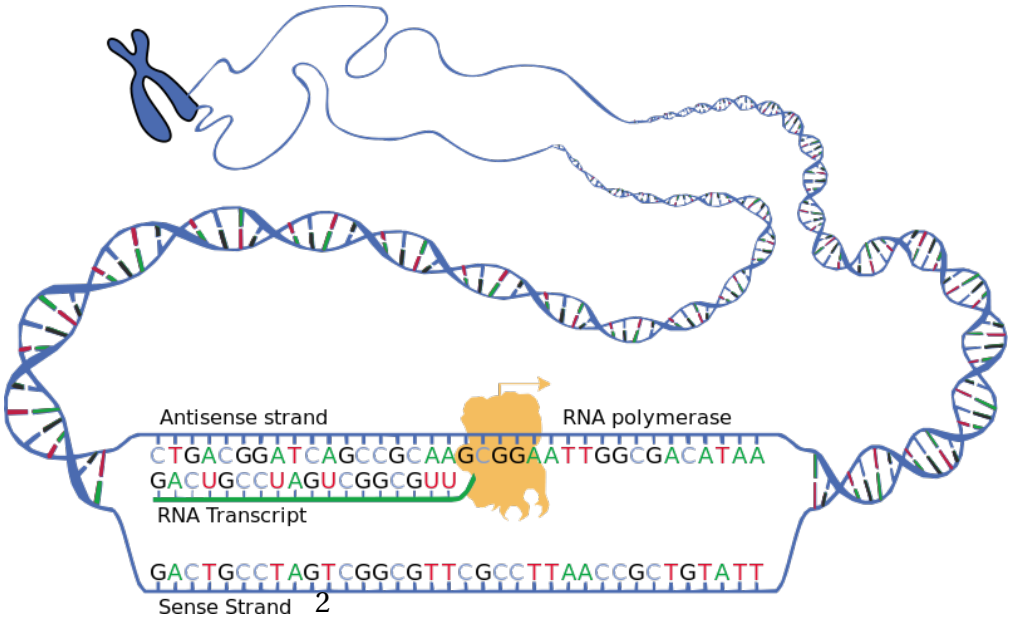
Promoter: Als Promoter wird der Teil der DNA bezeichnet, an dem eine codierende Sequenz (Exon) anfängt. Dieses ist meistens das Basentripel AUG, manchmal jedoch auch GUG oder GUU. Promoter können durch Proteine oder andere Stoffe in ihrer "reaktionsfreudigkeit"gehemmt oder angeregt werden, womit die Wahrscheinlichkeit auf eine Transkription natürlich wächst und fällt.

Transkriptionsfaktor: Protein oder Stoff im Zellkern, welches durch Wechselwirkung mit der DNA sich an diese anlagert und so den Transkriptionsvorgang einleitet.

Prozess in der Zelle

Der allgemeine Vorgang, wie aus denen in Chromosomen angeordneten DNA Strängen ein Protein entsteht ist, grob in drei Bereiche aufteilbar:

1. **Transkription:**



Ein Transkriptionsfaktor lagert sich an eine Promoterregion (CTG, TAA) an und bewirkt eine Elongation der Doppelhelix. Als codogener Strang (in der Abb. Antisense Strand) wird nun der Strang bezeichnet, an dem sich der Promoter befindet. Er ist auch der Strang der kopiert werden soll. An den freiliegenden Basen des codogenen Stranges lagern sich nun erneut die komplementären Ribonukleotide an. Der einzige Unterschied ist, dass statt Thymin nun Uracil als Bindungspartner verwendet wird. Dieser Vorgang stoppt bei dem abschließenden Terminator(UAG, UGA, UAA). Dieser angelagerte Strang wird RNA Transkript bezeichnet und löst sich nach dem Abschluss von dem codogenen Strang als prä-RNA.

2. Protein Prozessierung: Die in der Transkription entstandene prä-RNA kann nun durch verschiedene biologische Vorgänge noch erweitert werden. Durch das **Spleißen** werden möglicherweise noch vorkommende Introns aus der prä-RNA herausgeschnitten. Abschließend wird sie als mRNA ("messenger"RNA) aus dem Zellkern heraustransportiert.

3. Translation: Die mRNA enthält nun die Informationen die in der Zelle benötigt werden um das entsprechende Protein zu synthetisieren. Nach dem Verlassen des Zellkernes lagert sie sich an einem Ribosom an. Dieses Ribosom liest nun von dem Startcodon ausgehend die Basentripel ab und fügt sie mit einem jeweils passenden Basentripel (Aminosäuren) zusammen. Dies geschieht entlang des mRNA Stranges bis der Stopp Codon erreicht ist. Das durch die Verkettung von Aminosäuren entstandene Eiweiß löst sich vom Ribosom.

Das entstandene Produkt kann nun andere Gene bei der Proteinherstellung anregen oder hemmen. Das wird in einem Netzwerk dargestellt, welches später vorgestellt wird.

Beispiel:

Das kompliziert klingende, kann mathematisch etwas vereinfacht werden. Gen X erstellt mit der oberen Prozedur sein Enzym/Protein. Dieses hemmt Gen Y. Das bedeutet es wer-

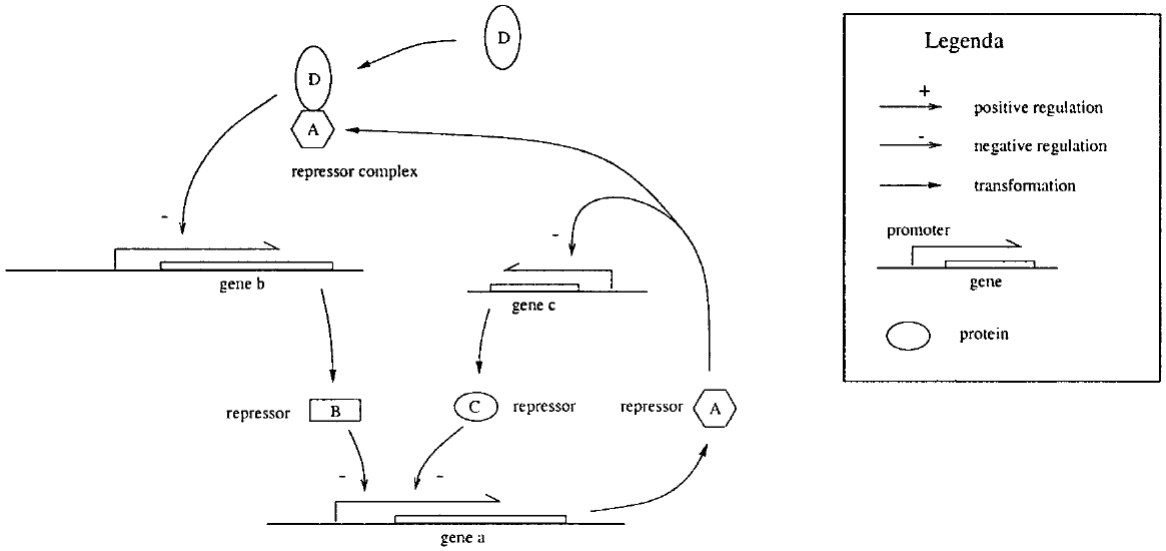
den weniger Transkripte von Gen Y erzeugt. Somit ist Gen Y weniger stark ausgedrückt und es sinkt in seinem Expressionslevel. Mathematisch auffällig sollte hier bereits sein, dass man dieses durch eine Markovkette modellieren könnte.

anschaulicheres Beispiel:

Sei unsere Zelle in der Bauchspeicheldrüse vorzufinden und unser Gen LacI sei dafür zuständig Laktose abzubauen. Laktose ist Milchzucker und wird von dem Körper als Alternative zu Glukose verwendet. Da die Glukosenutzung für den Körper deutlich einfacher ist bevorzugt er diese, solang Glukose vorhanden ist. Sofern nur Laktose vorliegt gibt es eine gewisse Konzentration an Laktosemolekülen. Durch die Laktosemoleküle wird im Zellkern nun ein Prozess in Gang gesetzt, sodass die Promoterregion für das LacI Gen angeregt wird. Das Gen stellt keine Laktoseabbau-Enzyme her, wenn keine Laktose vorliegt und wenn Glukose zusätzlich vorliegt. Es hat in unserem Modell später das Expressionslevel 0. Nun durch die Zugabe von Lactose steigt das Expressionslevel an, da durch anregen des Promoters auch Transkripte erzeugt werden. Es kommt eher zur Enzymherstellung. Sobald die Lactosekonzentration sinkt, wird die Anregung des Promoters wieder rückgängig gemacht und das Expressionslevel sinkt wieder. In der Realität ist der Vorgang natürlich um einiges komplizierter und tritt auch in veränderter Form auf. Fürs grobe Verständnis sollte es allerdings ein gutes Beispiel sein.

Mathematische Modelle

Im Folgenden betrachten wir wie wir Schritt für Schritt von einem GRN zu einem mathematischen Modell gelangen, welches uns dann am Ende die gewünschten Informationen über unser GRN ausgibt. In der Modellierung ist es häufig wichtig einen guten Mittelweg zu finden zwischen der Genauigkeit des mathematischen Modells und der Komplexität des Modells. Ziel ist es also, möglichst wenige Variablen und Rechnungen zu verwenden und dennoch ein der Realität sehr nahes Ergebnis zu erhalten. Nehmen wir als Beispiel folgende Situation an:



Was kann man aus diesem Netzwerk an Informationen herausbekommen?
Man kann relativ schnell erkennen, dass es sich um einen negativen Regulierungszyklus handelt. D zusammen mit dem von Gen A erzeugten Protein hemmen Gen B und Protein A hemmt seinerseits Gen C. Die Produkte von Gen C und B hemmen wiederum den Ausdruck von Gen A. So ergibt dies einen Regulierungszyklus, welcher eine niedrige Expression aller 3 Gene zur Folge hat.
Nun probieren wir einmal diesen Prozess mathematisch zu beschreiben.

Rate Equations

Es ist zunächst einmal naheliegend, ein Modell mit Hilfe der Konzentration der beteiligten Stoffe aufzubauen. Dies bedeutet im allgemeinen:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(\mathbf{x}) \qquad \mathbf{x} = [\, x_1, \ldots, x_n \,]$$

Die x_i stehen jeweils für die Konzentrationen verschiedener Stoffe. In die Funktion $f_i(\mathbf{x})$ kann nun alles gepackt werden, was die Konzentration von x_i beeinflusst. Das sind natürlich zum einen die Konzentrationen aller anderen Stoffe im System. Es könnte aber auch weitere äußere Faktoren geben, welche die Konzentration von x_i beeinflussen, wie zum Beispiel die Nahrungsaufnahme (Beispiel Laktose, welches die Erzeugung von Enzymen

zum Laktose abbau anregt).

Da für gewöhnlich in der Natur nichts geschieht ohne, dass eine gewisse Zeit vergeht, kann man der Zeit auch noch eine Bedeutung beimessen. Damit sähe die Formel dann folgendermaßen aus.

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_1(t-\tau_{i1}),...,x_n(t-\tau_{in}))$$

$\tau_1,...,\tau_n$ stehen in diesem Fall für die diskrete Zeitverzögerung zu der die Konzentration von x_1 eine Rolle spielte. Es kann genau so gut mittels Integralen eine kontinuierliche Zeitverzögerung genutzt werden.

Ein Geschwindikeitsgleichungssystem mit der nichtlinearen Regulierungsfunktion $r : \mathbb{R} \to \mathbb{R}$ und Konzentrationsvektor $\mathbf{x} = [\, x_1, \ldots, x_n \,]$ ist

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= \kappa_{1n}\mathrm{r}(x_n) - \gamma_1 x_1 \\ \frac{\mathrm{d}x_i}{\mathrm{d}t} &= \kappa_{i,i-1}x_{i-1} - \gamma_i x_i \end{aligned}$$

Interpretation dieses Systems ist:

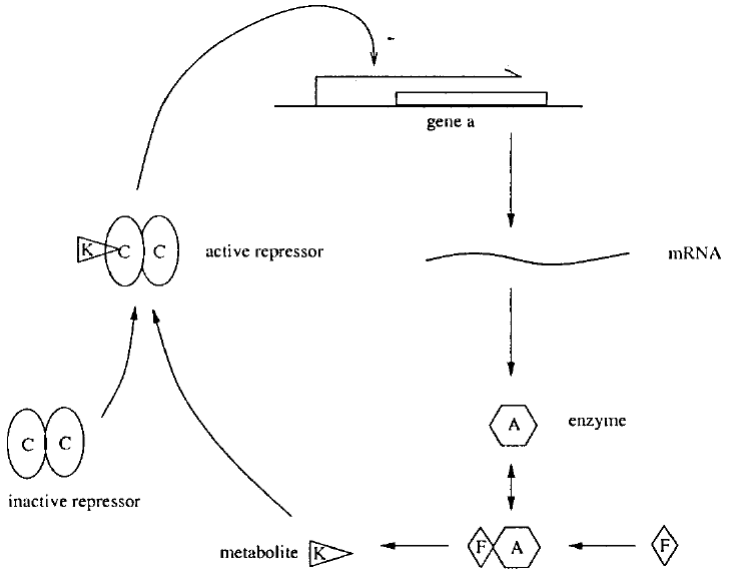
- $\kappa_{i,j}$ sind Produktionskonstanten. Sie multipliziert mit der Konzentration von x_{i-1} ergeben also den neu produzierten Teil des Stoffes x_i.
- γ_i sind Eliminationskonstanten. Sie können einen Zerfall oder eine Umwandlung des Stoffes repräsentieren.
- $\mathrm{r}(x_n)$ ist die Regulierungsfunktion. Sie steht für eine Funktion zwischen 0 und 1, welche den Zusammenhang zwischen der Produktion von x_1 mit der Konzentration von x_n verknüpft. Dies kann von Modell zu Modell unterschiedlich sein.

Bemerkung: Eine mögliche Regulierungsfunktion ist natürlich die Identität.

Beispiel eines RES:

Nutzen wir nun die Konzentration der mRNA als x_1, die des Proteins A als x_2 und die des Metabolites K als x_3. Dann ist folgendes das Rate Equationsystem zu dem rechts stehenden Netzwerk:

$$\begin{aligned}\dot{x}_1 &= \kappa_1 r(x_3) - \gamma_1 x_1 \\ \dot{x}_2 &= \kappa_2 x_1 - \gamma_2 x_2 \\ \dot{x}_3 &= \kappa_3 x_2 - \gamma_3 x_3\end{aligned}$$



$\kappa_1, \kappa_2, \kappa_3$ sind wiederum Produktionskonstanten und $\gamma_1, \gamma_2, \gamma_3$ sind Eliminationskonstanten. r(x) ist eine Regulierungsfunktion. Das Modell hilft uns die Konzentrationsveränderungen der einzelnen Stoffe zu jedem Zeitpunkt exakt zu beschreiben.

Gilt $\frac{dr}{dx_3} < 0$ bedeutet dies biologisch, dass es sich um einen negativen Produktionszyklus handelt. ("negative feedback loop") Dies ist dadurch bedingt, dass wenn unser Stoff x_3 (metabolit K) in der Konzentration ansteigt, so wird das Gen biologisch stärker gehemmt. Das bedeutet es wird weniger mRNA $= x_1$ produziert. Das ist der Grund für eine derartige Regulierungsfunktion.

Hill Curve

Für eine solche Regulationsfunktion stehen lineare, sigmoide, oder Stufenfunktionen zur Verfügung. Die Hill Kurve ist eine nicht lineare Regulationsfunktion definiert durch

$$H^+(x_j, \theta_j, m) = \frac{x_j^m}{x_j^m + \theta_j^m}.$$

Seien unsere werte x_j und θ_j größer als 0 und m > 0, so verläuft die Kurve zwischen 0 und 1 S-förmig.

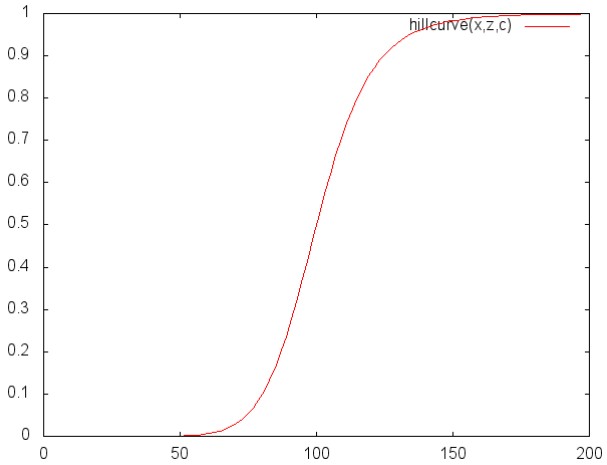


Abbildung 0.1: $\theta_j = 100$, m $= 10$

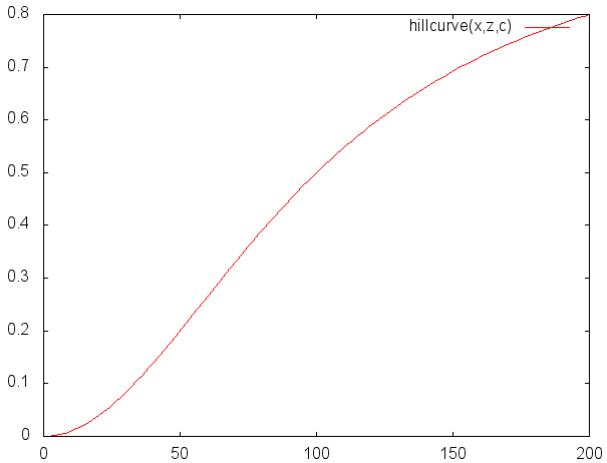


Abbildung 0.2: $\theta_j = 100$, m $= 2$

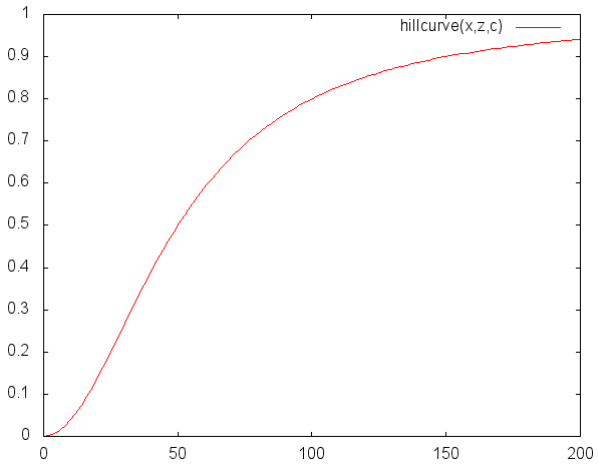


Abbildung 0.3: $\theta_j = 50$, m = 2

- Da θ_j eine Konstante ist, geht der Bruch bei großen x_j (im Vergleich zu θ_j) gegen 1, bei kleinen x_j gegen 0.
- Durch die Wahl von m und θ_j ist es möglich die Kurve wie in den 3 Abbildungen dargestellt zu verändern.
- Bei $x_j = \theta_j$ nimmt die Kurve den Wert 0.5 an.

Also für eine positive Regulierung benutzen wir die obenstehende Definition der Hillkurve. Für negative Regulierung verwenden wir folgende:

$$H^-(x_j, \theta_j, m) \quad = \quad 1 - H^+(x_j, \theta_j, m) \quad = \quad 1 - \frac{x_j^m}{x_j^m + \theta_j^m}.$$

Anwendung:

Für unser obiges Beispiel bedeutet dies folgendes:

$$\begin{aligned}
\dot{x}_1 &= \kappa_1 \left[1 - \frac{x_3^m}{x_3^m + \theta^m} \right] - \gamma_1 x_1 \\
\dot{x}_2 &= \kappa_2 x_1 - \gamma_2 x_2 \\
\dot{x}_3 &= \kappa_3 x_2 - \gamma_3 x_3
\end{aligned}$$

Mit der geeigneten Wahl von θ und m lässt sich das Modell nun modifizieren. Für große θ würde die Konzentration von x_1 schneller fallen. Interpretation dieser Gleichungen ist folgende:

- Die Konzentration an mRNA führt zu einer Produktion von Protein A. Gesamtprodukt von A: $\kappa_2 x_1$.
- Die Konzentration von A wird jedoch gleichzeitig vermindert, dadurch, dass es sich mit einem F verbindet. Eliminationsanteil: $\gamma_2 x_2$.
- Die Konzentration von Metabolit K steigt nun durch die Verbindung von A mit F um den Teil $\kappa_3 x_2$ an und vermindert sich durch das anlagern an die beiden CC um den Teil $\gamma_3 x_3$.
- Eine höhere Konzentration an Metabolit K führt zu mehr aktiven Repressoren und damit zu einer stärkeren negativen Regulierung von Gen a. Dies hat eine kleinere Produktion von mRNA zur Folge. Dies bewirkt die Regulierungsfunktion.

Der Term $\frac{x_3^m}{x_3^m + \theta^m}$ gibt uns also eine Art "wirkenden Konzentrationswertßwischen 0 und x_j aus, welcher für die mRNA Herstellung von Bedeutung ist.
Soll die Repression des Genes durch K erst spät einsetzen und auch schnell dazu führen, dass keine mRNA mehr produziert wird, so kann man dies über die geeignete Wahl der Parameter m und θ bestimmen.

Die RES können natürlich auch deutlich mehr Variablen als nur 3 beinhalten. Und damit auch Interaktionen zwischen verschiedenen Genen repräsentieren. Für den Einblick in die Methode der RES ist dies Beispiel jedoch ausreichend.

Von RES zu Wahrscheinlichkeitsmodellen:

Einblick:

Unsere RES geben uns zu jedem Zeitpunkt einen genauen Überblick über das Verhalten des Systems. Des Weiteren werden alle beteiligten Stoffe betrachtet und deren Konzentrationen. Es ist sozusagen eine mikroskopische Sicht auf ein Genregulationsmodell. Oft wird diese genau Sichtweise aber gar nicht benötigt und es ist lediglich das Langzeitverhalten und der Ausdruck eines Gens interessant.
Das bedeutet, dass der Verlauf der Proteinkonzentration nicht untersucht werden soll, sondern vielmehr das Produkt ganz am Ende. Natürlich ist es möglich das auch bei dem RES zu betrachten. Jedoch ist die Vielfalt an Variablen die man dafür verwendet oft sehr groß, welches das gesamte Modell sehr unübersichtlich und schwer zu berechnen macht.

In dem Artikel "Probabilistic representation of gene regulatory networks"von Linyong Mao und Haluk Resat wird daher ein Modell vorgestellt, welches den soeben vorgestellten Prozess mithilfe eines Wahrscheinlichkeitsmodells mathematisch modellieren soll um eine Vorhersage über die Expression eines bestimmten Genes machen zu können. Es soll eine Aussage über die Durchschnittsexpression von "Green Flourescent Protein"(GFP) gemacht werden können. Wir ordnen dafür zunächst jedem Gen ein bestimmtes Expressionslevel zu. Für uns sind diese möglichen Expressionslevel ganze Zahlen zwischen 0 und 200. Ein Expressionslevel steht dafür, wie stark das Gen ausgedrückt ist. (Wie viel Genprodukt erzeugt wird)
Mathematisch heißt das erstmal:

$$X_t \in \{0, ..., 200\} \qquad \forall t \in N$$

Als sinnvolle Notation bietet sich noch an einen weiteren Index k zu vergeben, welcher angibt, um welches Gen es sich handelt.

$$X_{k,t} \in \{0, ..., 200\} \qquad \forall t \in N$$

Ein einfaches Regulationsnetzwerk zwischen 3 Genen ist nun durch folgende Abbildung gegeben:

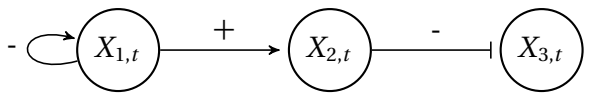
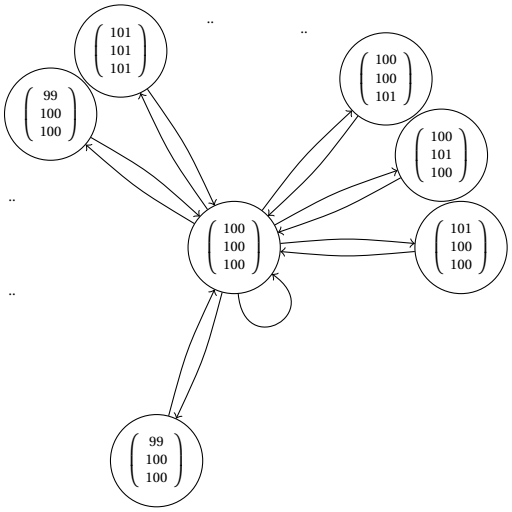


Abbildung: Genregulationsnetzwerk dreier Gene.

Die Pfeile verdeutlichen jeweils eine Regulation in Pfeilrichtung und mit dem Nebenstehenden Vorzeichen. Wie bereits erwähnt lassen wir den genauen Blick in das System außen vor und interessieren uns nur noch über die Interaktionen zwischen den verschiedenen Genen. Eine Markovkette die dieses modelliert sieht folgendermaßen aus, wobei die Übergangswahrscheinlichkeiten welche an den jeweiligen Pfeilen stünden erst gleich genauer untersucht werden. Der Zustandsraum besteht nun aus all den Vektoren welche aus den

Kombinationen der Expressionslevel der n Gene möglich sind. Dieser bleibt somit weiterhin endlich. Für unsere 3 Gene aus der Abbildung oben heißt das:



Regulationsstärken

Gesucht sind nun die an den Pfeilen stehenden Wahrschienlichkeiten. Wobei diese für jedes Gen einzeln geltenden Wahrschienlichkeiten dann zu der gesamten Wahrschienlichkeit jedes Schrittes führen. Für Gen k ist gesucht:

$P_{k,t}(\uparrow)$ =*Wahrscheinlichkeit dass Gen k im Zeitpunkt t+1 ein Expressionslevel steigt*
$P_{k,t}(\downarrow)$ =*Wahrscheinlichkeit dass Gen k zum Zeitpunkt t+1 ein Expressionslevel sinkt*
$P_{k,t}(-)$ =*Wahrscheinlichkeit dass Gen k im Zeitpunkt t+1 das gleiche Expressionslevel hat*

Um hier eine Wahrscheinlichkeit zu definieren die halbwegs Sinn macht, haben wir zunächst einmal einen Gewichtsfaktor eingeführt, der die Stärke der Regulationen durch verschiedene Produkte repräsentieren soll. Da die Regulation durch andere Gene bewirkt wird, benutzen wir zunächst einmal folgende Notation

$$W_{XY} = \text{Stärke der Regulation von Gen Y durch Gen X}$$
$$E_Y(t) = \text{Genexpressionslevel von Gen Y zum Zeitschritt t}$$

Da die Expression von Gen X für die Regulation von Gen Y natürlich auch eine Rolle spielt definiert man

$$S_Y(t) = \sum_{x \in \{\textit{Alle Gene ausgenommen Y}\}} W_{xY} * E_x(t),$$

als die Gesamtregulationsstärke von Gen Y.

Bemerkung 1: W_{XY} kann positiv oder negativ sein, was für Gen Y einer Anregung oder einer Hemmung entspricht.

Bemerkung 2: W_{XY} hängt nur von den korrespondierenden Genen, nicht von der Zeit oder der Expression der Gene ab.

Bemerkung 3: Offensichtlich kann auch $S_Y(t)$ positiv oder negativ sein.

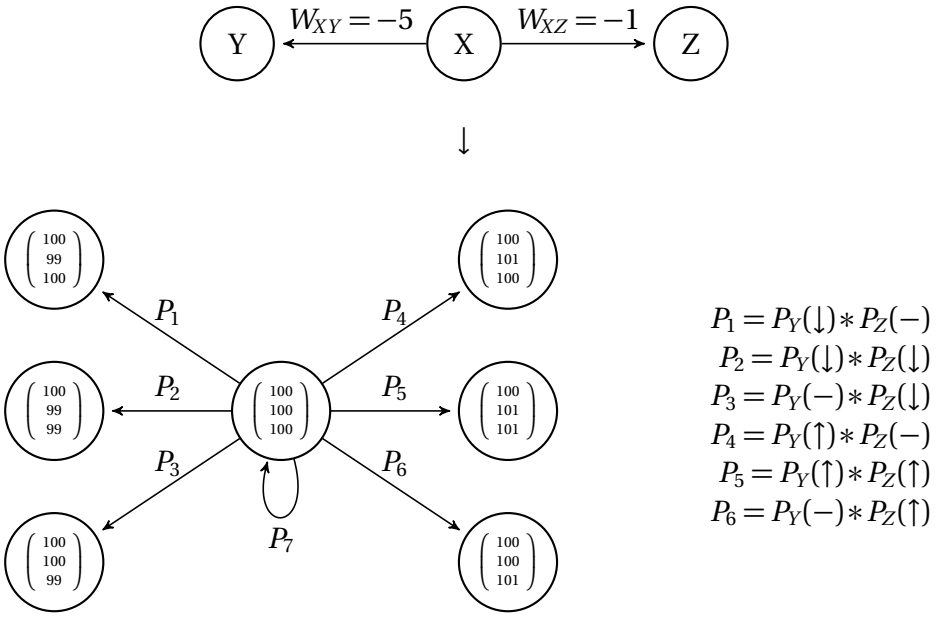
Für das Modell sind also folgende Implikationen sinnvoll:

$$\begin{aligned} S_X(i) > 0 \quad &\Longrightarrow \quad P_{X,i}(\uparrow) > P_{X,i}(\downarrow) \\ S_X(i) < 0 \quad &\Longrightarrow \quad P_{X,i}(\uparrow) < P_{X,i}(\downarrow) \end{aligned}$$

Anwendung der Hillkurve im WM:

Das folgende Beispiel beschreibt eine weitere Eigenschaft, welche unser Modell haben sollte.

kleines Beispiel:
Sehe das Regulationsnetzwerk und unser Markovmodell für den 1. Zeitschritt folgendermaßen aus:



$$\begin{aligned} P_1 &= P_Y(\downarrow) * P_Z(-) \\ P_2 &= P_Y(\downarrow) * P_Z(\downarrow) \\ P_3 &= P_Y(-) * P_Z(\downarrow) \\ P_4 &= P_Y(\uparrow) * P_Z(-) \\ P_5 &= P_Y(\uparrow) * P_Z(\uparrow) \\ P_6 &= P_Y(-) * P_Z(\uparrow) \end{aligned}$$

Abbildung: Markovmodell nur für 1. Schritt.

Sei $W_{XY} = -5$, $W_{ZY} = -1$ und die Anfangsexpression von Gen X, Gen Y und Gen Z seien 100. Gen X verändere sich nicht in der Expression, wie im Bild deutlich und übt lediglich einen Effekt auf Y und Z aus. Dieser ist unterschiedlich gewichtet. Für beide Gene gilt:

$$S_Y(t) = S_Y = -500 \longrightarrow P_{Y,t}(\uparrow) < P_{Y,t}(\downarrow)$$
$$S_Z(t) = S_Z = -100 \longrightarrow P_{Z,t}(\uparrow) < P_{Z,t}(\downarrow)$$

Gen Z soll natürlich nun eine geringere Wahrscheinlichkeit für den Sprung in ein niedrigeres Expressionslevel haben als Gen Y, da die Regulationsstärke auch deutlich kleiner ist. Da dieser Zusammenhang oft nicht linear ist brauchen wir also wieder eine Regulierungsfunktion , welcher den experimentellen Werten möglichst nahe kommt. Für den Übergang in einen weiteren Zustand betrachten wir nun exemplarisch einmal den Pfad mit P_2.
Als Wahrscheinlichkeit ergibt sich hierfür:

$$P_2 = P_Y(\downarrow) * P_Z(\downarrow)$$

Eigentlich müsste noch die Wahrschienlichkeit für $P_X(-)$ multipliziert werden. Diese soll aber einfachheitshalber 1 sein. Was dies Beispiel deutlich machen soll ist , dass folgende Implikation zusätzlich gelten sollte: $P_Y(\downarrow) > P_Z(\downarrow)$
Dies sollte erfüllt werden, da dann die stärkere Gesamtregulation von -500 auf Gen Y sich in der Wahrscheinlichkeit für eine negative Regulierung wiederfindet.

$\Rightarrow$ Idee: Benutze die Hillkurve und verwende $|S_Y(t)| = x_j$ um so die Wahrscheinlichkeit in Abhängigkeit von der Regulationsstärke auszudrücken.

Modifizierung

Um die Wahrscheinlichkeiten genauer zu beschreiben gehen nun die Informationen ein, die man aus Experimenten gewinnt. Anhand dieser experimenteller Daten wurde herausgefunden, dass mit der Hillkurve als Regulationsfunktion für die Wahrschienlichkeiten das beste Ergebnis in der Simulation erzielt wurde. Wobei in dieser nicht mehr die Konzentration eine Rolle spielt sondern vielmehr das Gesamtgewicht S_k.

$$H(S_k(t), C_k, n_k) = \frac{|S_k(t)|^{n_k}}{|S_k(t)|^{n_k} + C_k^{n_k}}$$

Hier sind C_k und n_k für jedes Gen unterschiedliche Konstanten, welche man wiederum durch den Abgleich mit experimentellen Daten passend gewählt hat. Veränderungen von C_k und n_k haben hier die gleichen Auswirkungen, welche bereits bei der Einführung der Hillkurve beschrieben wurden.

Grundsätzlich gibt es in unserem Modell eine Wahrscheinlichkeit P_0 mit welcher ein Gen in das nächst höhere oder niedrigere Expressionslevel springt.

Frage: Wenn ein Gen negative Regulation erfährt, warum kann es dann dennoch im Expressionslevel steigen?

Das liegt daran, weil eine negative Regulation nicht gleichzeitig bedeutet, dass das Gen bis zu einer Expression von 0 reguliert wird. Würden wir ein Modell bauen, indem es dann unmöglich wäre für das Gen ein Level höher zu gelangen, so hätten wir eine Kette die gegen das Expressionslevel 0 streben würde, da es nur die beiden Möglichkeiten gibt in einem Zustand zu verweilen oder aber den nächstniedrigeren zu erlangen.
Die Bestimmung von P_0 erfolgt wiederum durch den Abgleich mit den experimentellen Werten. Abhängig davon ob ein Gen nun angeregt oder gehemmt wird ergeben sich die folgenden Wahrscheinlichkeiten. Die linke Seite entspricht dem Fall eines gehemmten Gens und die rechte Seite dem eines angeregten.

$S_x(t) < 0$	$S_x(t) > 0$
$P_{x,t}(\uparrow) = P_0$	$P_{x,t}(\uparrow) = P_0 * \left[1 + \frac{\lvert S_k(t)\rvert^{n_k}}{\lvert S_k(t)\rvert^{n_k} + C_k^{n_k}}\right]$
$P_{x,t}(\downarrow) = P_0 * \left[1 + \frac{\lvert S_k(t)\rvert^{n_k}}{\lvert S_k(t)\rvert^{n_k} + C_k^{n_k}}\right]$	$P_{x,t}(\downarrow) = P_0$
$P_{x,t}(-) = 1 - P_{x,t}(\downarrow) - P_{x,t}(\uparrow)$	$P_{x,t}(-) = 1 - P_{x,t}(\downarrow) - P_{x,t}(\uparrow)$

Einige Interpretationen dieser Wahrscheinlichkeiten (In unserem Modell wird der Wert $P_0 = 0.2$ verwendet):

- Der Ausdruck $\left[1 + \frac{\lvert S_k(t)\rvert^{n_k}}{\lvert S_k(t)\rvert^{n_k} + C_k^{n_k}}\right]$ liegt auf jedenfall zwischen 1 und 2.
- Damit wiederum liegt die Wahrscheinlichkeit für $P_{x,t}(-)$ zwischen 0.6 und 0.4.
- Bei einer starken positiven Regulation ist die Wahrscheinlichkeit von $P_{x,t}(\uparrow)$ fast doppelt so groß wie die von $P_{x,t}(\downarrow)$.

Natürlich kann kein Expressionslevel kleiner als 0 und keines größer als 200 erreicht werden. Die Wahrscheinlichkeiten für diese Fälle dürften klar sein.

Man sieht auch direkt, dass die Eigenschaften für eine Markovkette erfüllt sind. Die Wahrscheinlichkeit ist ausschließlich durch das Expressionslevel vom vorherigen Zeitpunkt bestimmt. Die davor angenommenen Level spielen keine Rolle.

Zeige nun: Die Markovkette ist ergodisch.

Hierfür definieren wir zunächst den endlichen Zustandsraum:

$$\mathbf{Z} = \{\, \mathrm{X} = (\, X_1, \ldots, X_n \,) \mid \text{mit } X_i \in \{0, \ldots, 200\} \,\}$$

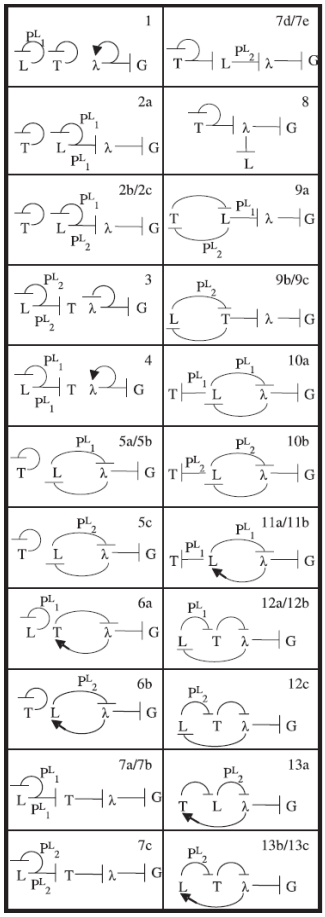
Zeigen wir zunächst die Homogenität:

homogen: Die Abbildung im Beispiel gibt eine gute Anschauung, warum die Wahrscheinlichkeiten $P_{x,t}(\uparrow)$, $P_{x,t}(\downarrow)$ und $P_{x,t}(-)$ nicht von t abhängen sondern lediglich von dem Zustand in dem Sie zum vorherigen Zeitpunkt waren. Das t in dem Ausdruck $S_k(t)$ sollte nur deutlich machen, dass dies kein festes Gewicht war, sondern eines, welches in jedem Zeitschritt von den Expressionsleveln abhängt. Somit haben wir eine homogene Markovkette.

Da $P_{x,t}(\uparrow)$, $P_{x,t}(\downarrow)$ und $P_{x,t}(-)$ in unserem Modell alle größer als 0 sind für alle Gene x, folgt bereits die **Irreduzibilität**.
Die **Aperiodizität** ist auch leicht erkennbar, wenn man den beliebigen Startzustand a_i (Vektor von Expressionsleveln) betrachtet. Und die mögliche Anzahl an Schritten untersucht, nach denen es mit positiver Wahrscheinlichkeit möglich ist hierher zurückzugelangen, so stellt man fest, dass es mit jeder Schrittzahl möglich ist.
Unsere Markovkette ist somit ergodisch und somit gibt es einen eindeutigen stationären Zustand.

Exemplarische Anwendung

Ziel: Wir betrachten das unten stehende System von 22 verschiedenen Genregulationsnetzwerken vierer Gene und wollen wie schon in der Einleitung dieses Abschnittes beschrieben, das Expressionslevel von GFP ermitteln können. Wir simulieren mit Hilfe des Computers den Ausdruck der Gene für die verschiedenen Zeitschritte in einem Regulationsnetzwerk. Dabei wird festgehalten wann sich ein Gen in welchem Expressionslevel befindet. Am Ende schauen wir welches Expressionslevel am Häufigsten angenommen wurde. Dieses ist unser Hauptexpressionslevel.
Die in dem Modell beteiligten Gene sind folgende.



Gene:

$\mathbf{G} = \text{GFP}$
$\mathbf{L} = \text{LacI}$
$\lambda = \lambda\text{-cI}$
$\mathbf{T} = \text{tetR}$

Jedes Gen hat ein bestimmtes Expressionslevel. Wir definieren

$$\mathbf{G}_t, \mathbf{T}_t, \lambda_t, \mathbf{L}_t \in \{0, ..., 200\} \qquad \forall\, t \in N$$

Hierbei entspricht t dem diskreten Zeitschritt und der Wert der angenommen wird dem Expressionslevel des jeweiligen Gens zum Zeitschritt t.
Am Einfachsten erscheinen hier natürlich die Netzwerke, die wenig Verbindungen und somit wenig Regulation beinhalten. Die nötigen Werte für die Gewichte und den Konstanten C_k und n_k erhält man durch den Abgleich mit den experimentellen Werten und der schrittweisen Verbesserung.

Für das Modell, welches in dem Artikel beschrieben wurde, werden folgende Werte verwendet:

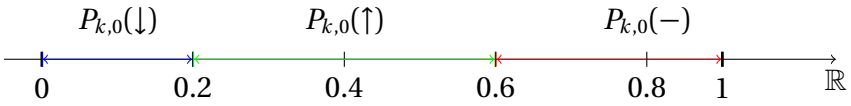
W(λ,+)	=	1.0
W(λ,-)	=	-1.0
W(L,P_1^L)	=	-4.0
W(L,P_2^L)	=	-1.0
W(T,-)	=	-4.0

Gen	C_k	n_k
G	100	2.0
λ	150	3.0
L	100	2.0
T	200	1.5

Doch wie berechnen wir nun das Hauptexpressionslevel?

Erläuterung der Berechnung des Hauptexpressionslevels

Wir benutzen eine willkürliche Grundkonfiguration an Expressionsleveln für die vier verschiedene Gene. Hiermit lässt sich nun $S_k(0)$ berechnen für alle k $\in$ {G,L,T,λ }. Wir lassen uns vom Computer eine Zufallszahl zwischen 0 und 1 ausgeben und teilen das Einheitsintervall in drei Abschnitte. Das Gen springt dann folgendermaßen in das neue Expressionslevel. Landet unsere Zufallszahl in dem Intervall zwischen 0 und $P(\downarrow)$ so sinkt das Expressionslevel. Zwischen $P(\downarrow)$ und $P(\downarrow)+P(\uparrow)$ so steigt das Expressionslevel und zwischen $P(\downarrow)+P(\uparrow)$ und 1 bleibt es gleich.



Iterativ führen wir dies für alle vier Gene zum Zeitschritt 0 durch und erhalten so eine neue Verteilung der Genexpressionslevel zum Zeitpunkt t=1. Hier berechnen wir wieder $S_k(1)$ für alle k $\in$ {G,L,T,λ} und lassen uns für jedes Gen erneut eine Zufallszahl ausgeben und erhalten die neuen Expressionslevel aller Gene.

Insgesamt lassen wir den Computer 60 Millionen solcher Schritte berechnen. Wir speichern alle Schwankungen in den Expressionsleveln ab und berechnen nachher das Expressionslevel, in welchem sich das Gen am häufigsten befand. Für genaue Werte sorgt bereits eine Länge von 15 Millionen Schritten. Wir führen 6 Simulationen mit 60 Millionen Schritten durch und benutzen den Durchschnittswert/ am häufigsten angenommenen Wert. Der Wert sollte also sehr präzise sein.

Anwendungsbeispiel:

Für jedes dieser 22 Netzwerke wurde mit dem Computer das Hauptexpressionslevel vom GFP berechnet und mit dem Wert der durch Experimente bestimmt wurde verglichen.
Einige Experimente lieferten 2 Ergebnisse. Das kam durch Mutationen in der Regulation zu Stande. Diese wurden in der Analyse nachher nicht berücksichtigt.
Rückwirkend haben wir die experimentellen Daten schon vorher gehabt. und unsere Simulation können wir nun mit verschiedensten Parametern für die Gewichte und die C_k und n_k häufiger laufen lassen und die Ergebnisse mit den Werten aus dem Experiment vergleichen. So ist es möglich das Ergebnis immer genauer zu machen.

Network ID	GFP (expt)	GFP (sim)	LacI (sim)	TetR (sim)	λ-cI (sim)
1	2	1	6	8	199
2a	2	13	6	8	57
2b/2c	60/29	6	7	8	99
3	13	17	16	40	30
4	0	1	6	21	199
5a/5b	20/80	70	79	8	22
5c	99	12	15	8	90
6a	1	92	6	106	2
6b	90	80	128	8	5
7a/7b	28/36	39	6	21	15
7c	65	58	16	40	9
7d/7e	1/1	7	11	8	93
8	29	19	19	8	37
9a	0	8	2	97	87
9b/9c	12/98	85	2	97	3
10a	94	70	81	11	21
10b	0	13	16	72	91
11a/11b	73/95	94	104	2	2
12a/12b	23/28	24	21	10	23
12c	16	39	31	22	14
13a	0	6	1	192	100
13b/13c	1/0	7	193	2	96

mögliche Erweiterungen:

Das Modell kann noch in mehrere Richtungen verfeinert werden. Beispielsweise könnte man:

- Man könnte einen chemischen Inducer einführen, welcher die reaktionsfreudigkeit eines Genes zusätzlich hemmt oder anregt.
- Man kann ein Modell aus noch mehr verschiedenen Genen erzeugen.

Was haben wir gelernt?

Insgesamt sollte der Vortrag einen Bezug zwischen der Mathematik und Realität vermittelt haben. Er stellt zwei verschiedene Methoden vor ein Gen Regulationsnetzwerk zu beschreiben, wobei der Hauptteil sich mit dem stochastischen Modell beschäftigt. Er sollte einen guten Überblick geben über die Modellierung eines GRN.