Zusammenfassung Systembiologie - FS18

v1.1

Madlaina Mettier, Gleb Ebert

29. Juli 2018

Vorwort

Diese Zusammenfassung enthählt wenn verfügbar alle Lernziele und passende Beispielaufgaben der Vorlesung Systembiologie (Stand FS18). Lernziele bzw. Aufgaben sind jeweils in **fett** geschrieben. Antworten bzw. Lösungen folgen direkt danach in normaler Schrift. Bei fehlenden Lernzielen wurden die wichtigsten Konzepte und Beispiele aus den Vorlesungsfolien zusammengefasst. Die Nummern der Kapitel entsprechen den Vorlesungswochen. Kapitel 5 entspricht dabei den Wochen 5 und 6. Wir können leider weder Vollständigkeit noch die Abwesenheit von Fehlern garantieren. Insbesondere unsere Lösungen der Aufgaben wurden nach besten Wissen und Gewissen erstellt und müssen nicht in dieser Form korrekt sein. Für Fragen, Anregungen oder Verbesserungsvorschlägen können wir unter `glebert@student.ethz.ch` erreicht werden. Die neuste Version dieser Zusammenfassung kann stets unter `https://n.ethz.ch/~glebert/` gefunden werden.

Inhaltsverzeichnis

1 Einführung

1.1 Lernziele

Limitationen des traditionellen molekularbiologischen Ansatzes anhand von Beispielen aufzeigen. Vom Human Genome Project hatte man sich erhofft, allen Genen eine Funktion zuzuordnen zu können. Es gibt aber viel mehr zelluläre Funktionen als Gene. Um das System zu verstehen, muss man die einzelnen Komponenten, ihre Interaktionen und wie eine Veränderung sich auf das System auswirkt verstehen.

Den Ansatz und die Ziele der Systembiologie beschreiben und Beispiele für Fragestellungen nennen, mit denen sich die Systembiologie beschaftigt. In der Systembiologie versucht man mithilfe computergestützter mathematischer Modelle zu verstehen, wie das funktionierende Zusammenwirken der molekularen Komponenten zu dem Verhalten und den Eigenschaften führt, die biologische Systeme zeigen.

Sie können Faktoren nennen, die zur Komplexität biologischer Systeme beitragen. Die Komplexität wird schnell zu hoch wenn verschieden Prozesse und Regelmechanismen wie Genexpression, Signalverarbeitung und Rückkopplung ineinander greifen. Dabei hat man oft eine fast unendliche Menge Daten und Möglichkeiten diese zu verknüpfen.

An Beispielen aufzeigen, warum Modelle nützlich sind, um biologische Fragestellungen zu beantworten, die man nicht allein durch Intuition lösen kann. An einem Beispiel erklären, wie man bei einer Systemanalyse grundsätzlich vorgeht. Mögliche Fragestellungen sind die Systemanalyse eines genregulatorischen Modells oder Michaelis-Menten-Kinetik. Misst man z.B. die Expression eines Gens in Populationen mit unterschiedlich grossen Aminosäurevorräten, muss bedacht werden, dass Zellen mit mehr AS grösser wachsen könnten und die Expressionsprodukte so verdünnter wären.

Ausgehend von einem Schema eines Reaktions- oder Pathway-Systems, die Geschwindigkeitsgesetze aufstellen, die dieses System beschreiben (Vertiefung in Woche 2).

$$A + B \xrightarrow{k} C \qquad \frac{d[A]}{dt} = -k[A][B]$$
$$\frac{d[B]}{dt} = -k[A][B] \qquad \frac{d[C]}{dt} = +k[A][B]$$

Die Annahmen nennen, die beim Massenwirkungsgesetz getroffen werden. konstante Temperatur, kein Katalysator

Ausgehend von den Elementarreaktionen einer Enzym-katalysierten Reaktion die Michaelis-Menten-Gleichung und ähnliche Gleichungen einfacher Enzymkinetiken herleiten und angeben, welche Annahmen man dabei trifft. Vereinfachungen: Produkt wird in irreversibler Reaktion hergestellt; $[E]_{\text{total}} = [E] + [E * S]$

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E * S \xrightarrow{k_2} E + P$$
$$\frac{d[S]}{dt} = -k_1[E][S] + k_{-1}[E * S]$$
$$\dots$$
$$v = \frac{v_{max}[S]}{K_m + [S]} \qquad \text{mit} \qquad v_{max} = k_{cat}[E]$$
$$K_m = \frac{k{-}1 + k_2}{k_1}$$

Annahmen: *[ES]* ist konstant; Enzyme sind gesättigt mit Substrat

Für jeden Parameter der Michaelis-Menten-Gleichung beschreiben, wof¨ur dieser inhaltlich steht und wie sich die der Kurvenverlauf qualitativ verändert, wenn dieser Parameter verändert wird. v_{max}: maximale Reaktionsrate; K_M: Affinität des Enzym-Substrat-Komplexes (Substratkonzentration bei 50% v_{max})

Andere Beispiele neben der Enzymkinetik nennen, auf die die Michaelis-Menten-Gleichung übertragbar ist und wissen die Parameter der Gleichung inhaltlich zuzuordnen. Bindung eines Trankriptionsfaktors T an eine DNA-Sequenz G: $[T]$ entspricht der Substratkonzentration; K gibt Affinität von T an G an

$$[G * T] = \frac{[G]^T * [T]}{[T] + K}$$

1.2 Beispielaufgaben

1.2.1

Die irreversible Reaktion $S \to P$ wird vom Enzym E mit dem katalytischen Koeffizienten k_{cat} (1 s^{-1}) und der Substrataffinität K_M (1 mM) katalysiert. Die Reaktion folgt einer Michaelis-Menten-Kinetik.

(a) Welche Gleichung beschreibt die Reaktionsrate v als Funktion der Substrat- $[S]$ und Enzymkonzentration $[E]$?

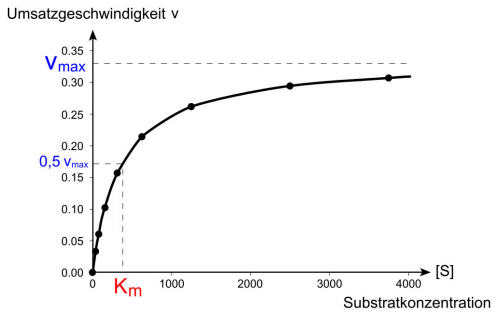
(b) Berechnen Sie die Reaktionsrate (in mM s^{-1}) für die Konzentrationen $[E] = 3$ mM und $[S] = 2$ mM.

(c) Zeichnen sie schematisch einen v-s-Plot mit dem Kurvenverlauf und geben Sie die Position von K_m und v_{max} an.

(a) $v = \frac{v_{max}[S]}{K_m+[S]} = \frac{k_{cat}[E][S]}{K_m+[S]}$

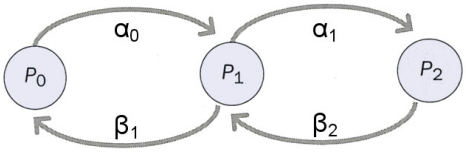
(b) $v = \frac{1s^{-1}*3mM*2mM}{1mM+2mM} = \frac{6mM^2s^{-1}}{3mM} = 2mMs^{-1}$

(c)



1.2.2

Formulieren Sie die Geschwindigkeitsgesetze für das Proteinsystem in dem Signalweg, der in der folgenden Abbildung dargestellt ist. Tipp: Modellieren Sie jeden Prozess als das Produkt einer Geschwindigkeitskonstante und der Konzentration der beteiligten Variable(n).

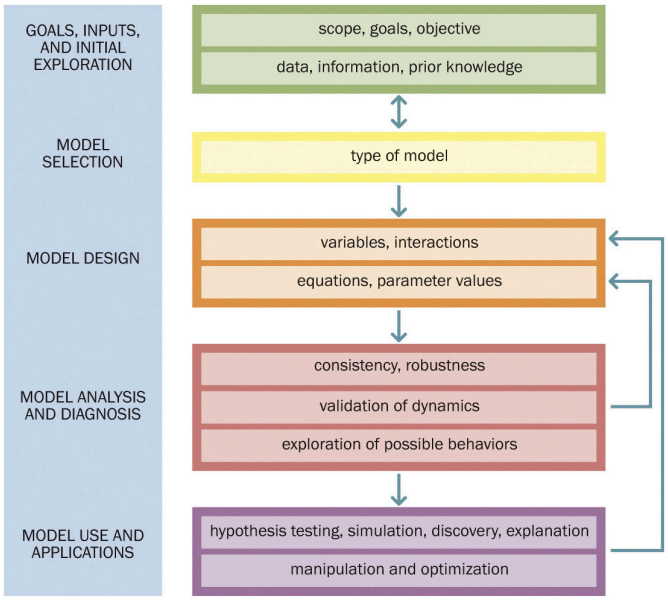


$$\frac{d[P_0]}{dt} = -k_{\alpha_0}[P_1] + k_{\beta_1}[P_1] = (k_{\beta_1} - k_{\alpha_0})[P_1]$$
$$\frac{d[P_1]}{dt} = -k_{\alpha_1}[P_2] - k_{\beta_1}[P_0] + k_{\alpha_0}[P_1] + k_{\beta_2}[P_2]$$
$$\frac{d[P_2]}{dt} = -k_{\beta_2}[P_2] + -k_{\alpha_1}[P_1]$$

2 Modellierung von Enzymreaktionen

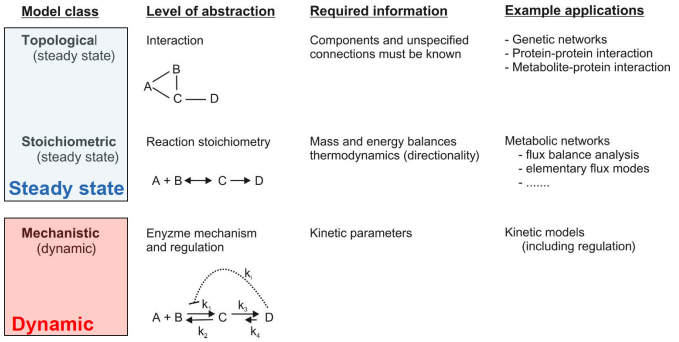
2.1 Lernziele

Den allgemeinen Arbeitsablauf beim Aufstellen eines Modells beschreiben.



Die Rolle der Auswahl einer angemessenen Zeit- und Grössenskala bei dem Aufstellen eines Modells erklären. Man muss die Zeit und Grösse das Modellorganismus oder Modelles allgemein der Fragestellung anpassen. Z.B. sind für geologische Prozesse Jahrhunderte nötig während für Diffusion in einer Zelle Sekunden ausreichen.

Die Unterschiede zwischen dynamischen und statischen Modellen erklären und für eine gegebene Fragestellung eine geeignete Modellwahl treffen. Dynamische Modelle verwendet man, wenn zeitliche Veränderungen bei dem modellierten Prozess eine Rolle spielen. Für die Beantwortung zeitunabhängiger Fragestellungen eignen sich statische Modelle.



Den Begriff steady state definieren und biologische Beispiele für steady-state-Bedingungen nennen. Das System befindet sich im GGW-Zustand, wenn sich die Konzentrationen der Metaboliten nicht verändern. Die Ein- und Ausflüsse eines Metaboliten halten sich die Waage.

Die Vorteile der Annahme eines *steady states* bei der mathematischen Modellierung erklären. Er ist eine mathematische Vereinfachung, weil man durch die zeitliche Konstanz Gleichungssysteme statt Differentialgleichungen aufstellen kann.

Die Begriffe „Konstante“, „Parameter“, „Variable“, „Zustandsgrösse“ (*state variable*) und „Systemzustand“ (*system state*) definieren. Konstante: Gleichbleibende Zahl
Zustandsgrösse: Grössen die den Zustand biologischer Objekte wie Gene, Zellen oder Moleküle als auch das Verhalten des Modells beschreiben.
Variable: Wert veränderlich und hängt von anderen Komponenten des Modells ab.
Parameter: Durch abschätzen, Literatur oder Experimente festgelegte Werte (z.B. K_M).
Systemzustand: Der Zustand des Systems zu einem spezifischen Zeitpunkt t, zu dem alle Werte bekannt sind.

Die ODE-Modelle für die elementaren Reaktionen aufstellen, die man in einem biochemischen Netzwerk findet, und kennen deren Lösung für einzelne Reaktionsschritte.
Konstante Produktion

$$\xrightarrow{k_1} A$$
$$\frac{d[A]}{dt} = k_1 \Rightarrow [A] = [A]_0 + k_1 * t$$

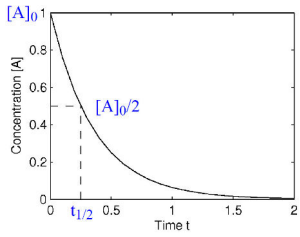
Linearer Abbau

$$A \xrightarrow{k_1}$$

$$\frac{d[A]}{dt} = -k_1[A] \Rightarrow \frac{d[A]}{[A]} = -k_1 * dt$$

$$\int_{[A]_0}^{[A]} \frac{d[A]}{[A]} = -k_1 \int_0^t dt = [A] = [A]_0 * e^{-k_1 * t}$$

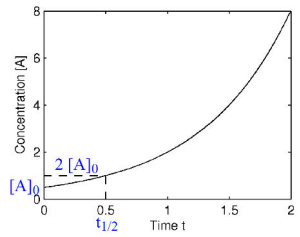
$$t_{1/2} = \frac{ln(2)}{k_1}$$



Autokatalyse

$$\xrightarrow{k_1 * [A]} A$$

$$[A] = [A]_0 * e^{k_1 * t} \qquad t_{1/2} = \frac{ln(2)}{k_1}$$



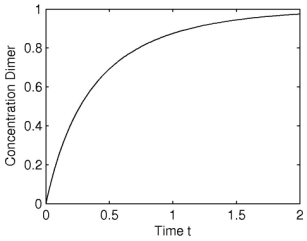
Dimerisierung

$$A + B \xrightarrow{k_1 * [A]} A * B$$

$$[A] = [A]_0 - [A * B] \qquad [B] = [B]_0 - [A * B]$$

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_1[A][B] \qquad \frac{d[A * B]}{dt} = k_1[A][B]$$

$$[A * B] = \frac{[A]_0[B]_0(1 - e^{-k_1 t([A]_0 - [B]_0)})}{[A]_0 - [B]_0 e^{-k_1 t([A]_0 - [B]_0)}}$$



Reversible Dimerisierung

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} A * B$$

$$[A * B]_{SS} = \frac{k_1}{k_{-1}} \qquad K_D = \frac{k_{-1}}{k_1}$$

$$\frac{d[A * B]}{dt} = +k_1[A][B] - k_{-1}[A * B] = 0$$

Produktion und Degradation

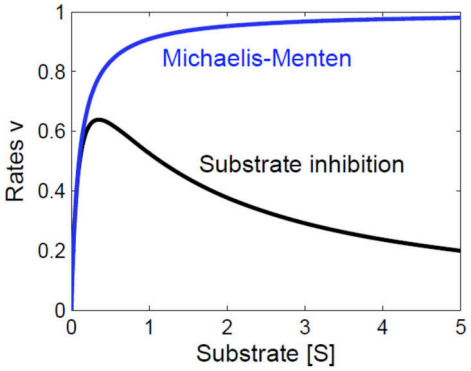
$$\xrightarrow{k_1 u} A \xrightarrow{k_2} \qquad [A]_{SS} = \frac{k_1 * u}{k_2}$$

$$\frac{d[A]}{dt} = +k_1 * u - k_2[A] \qquad [A] = \frac{k_1 * u}{k_2}\left(1 - e^{-k_2 * t}\right)$$

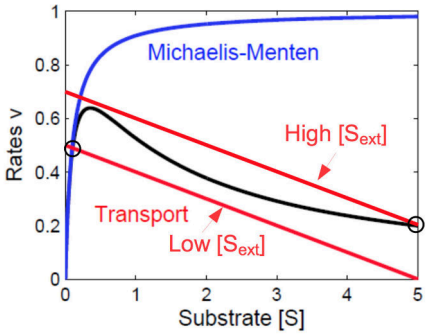
Ausgehend von einem Schema eines mehrschrittigen Reaktions- oder Pathway-Systems die Geschwindigkeitsgesetze aufstellen, die dieses System beschreiben. Die Konzentrationsveränderung über Zeit setzt sich immer aus allen Zu- und Abflüssen zusammen.

z.B. $\frac{d[A]}{dt} = k_1 * u - k_2[A]$ oder $[A]_{SS} = \frac{k_1 * u}{k_2}$

Erklären, wie sich die Reaktionsgeschwindigkeit in Abhängigkeit der Substratkonzentration zwischen Enzymen, die der Michaelis-Menten-Kinetik folgen, und Enzymen, die Substratinhibition zeigen, unterscheidet und die Kurvenverläufe qualitativ aufzeichnen.



Anhand einer Zeichnung erklären, wie sich die *steady-state*-Konzentration schalterartig mit Veränderung einer Kinetik (z.B. durch Änderung der externen Substratkonzentration) ändern kann. Wo befinden sich die steady states?



Substratinhibition	$v([S]) = \frac{v_{max}[S]}{[S] + K_M + \frac{[S]^2}{K_I}}$
Transport	$v_{\text{Trans.}}([S_{ext}], [S]) = D([S]_{ext} - [S])$

Der Schnittpunkt der schwarzen und roten Linie ist der *steady state*. Die schwarze Linie beschreibt die Substratinhibition.

2.2 Beispielaufgaben

2.2.1

Sie möchten die ersten 5 Minuten der Stoffwechselreaktion von Leberzellen auf die Zugabe von Inhibitoren der Glykolyse vorhersagen. Dazu entwickeln Sie ein Modell welches Sie mit der Reaktion auf einen bekannten Inhibitor trainieren.

(a) Welche Art von Modell brauchen Sie für diese Aufgabe?

(b) Welche Informationen und welche Daten brauchen Sie dafür?

(c) Beschreiben Sie kurz wie Sie das Problem angehen, wie Sie das Modell entwickeln werden, und was Ihr Vorgehen ist falls das Modell die Daten nicht beschreiben kann.

(a) dynamisch

(b) beteiligte Reaktionen, kinetische Parameter, Anfangskonzentrationen

(c) spezifische Frage, Modell auf Robustheit prüfen, Dynamik validieren, mögliche Verhalten vorhersagen, Modell implementieren, Hypothese testen
allfällige Abweichungen erklären und Modell ggf. anpassen

2.2.2

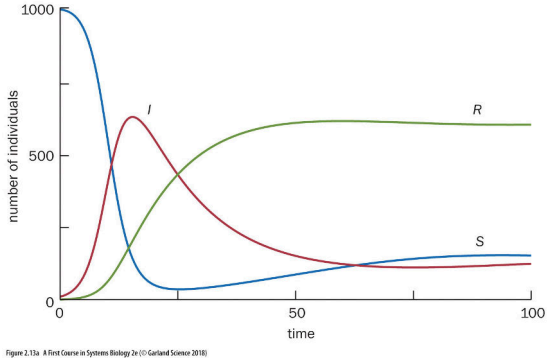
Ein Systembiologe möchte die Ausbreitung einer ansteckenden Krankheit modellieren, um die Anzahl von Individuen in einer Population abzuschätzen, die mit der Krankheit infiziert sind (I), empfänglich für eine Infektion sind (S) und resistent gegen die Krankheit sind, nachdem sie von ihr genesen sind (R). Er hat ein Modell aufgestellt, angemessene Werte für die Parameter angenommen und den Verlauf der Grössen I, S und R über die Zeit grafisch dargestellt.

(a) Hat der Systembiologe ein statisches oder dynamisches Modell gewählt? Begründen Sie ihre Antwort.

(b) Für welchen Zeitraum eignet sich die Annahme eines *steady states*? Markieren Sie diesen in der Zeichnung.

(c) Was ist der Vorteil einer *steady-state*-Annahme für die Modellierung eines Systems?

(d) Werden sich für die Variablen I, S und R, unabhängig von den Annahmewerten für die Modellparameter, immer die im Diagramm oben gezeigten *steady-state*-Werte einstellen? Begründen Sie Ihre Antwort.



(a) dynamisch, da Zeitdimension berücksichtigt wird

(b) ab ~ 75 auf Zeitskala

(c) mathematisch einfacher und damit weniger rechenintensiv

(d) nein, z.B. oszillierend wenn Resistenz nur kurz anhaltend

2.2.3

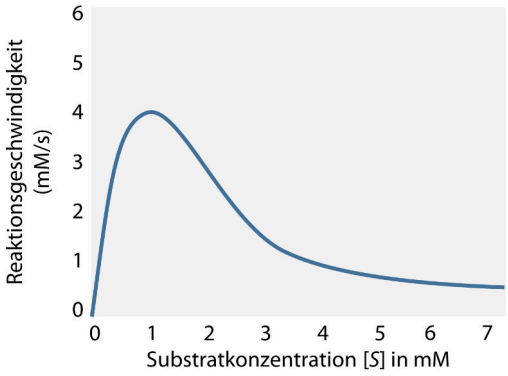
Ein Substrat S wird mit der Rate v_D von der Zelle über die Plasmamembran aufgenommen und in einem irreversiblen metabolischen Pathway weiterverwendet. Die Kinetik des ersten Enzyms des Pathways ist in der folgenden Abbildung gezeigt. Das System sei immer im *steady state*.

(a) Um welche Art von Kinetik handelt es sich bei dem Enzym?

(b) Die Aufnahmegeschwindigkeit in die Zelle v_D sei gegeben durch die Gleichung $v_D = D([S_{\text{extern}}] - [S])$, wobei $D = 1s^{-1}$ eine Diffusionskonstante und $[S_{\text{extern}}]$ die Substratkonzentration ausserhalb der Zelle ist. Zeichnen Sie v_D als Funktion von $[S]$, wenn $S_{\text{extern}} = 4mM$ in das Diagramm ein.

(c) Begründen Sie, warum die interne Substratkonzentration bei (b) im *steady state* $[S] < 1mM$ sein muss.

(d) Schätzen Sie ab, welche interne Substratkonzentration sich für $[S]_{\text{extern}} = 6mM$ einstellen wird. Zeigen Sie auch dies zeichnerisch.



(a) Substratinhibition

(b) $v_D = D([S_{\text{extern}}] - [S]) = D[S_{\text{extern}}] - D[S]$
$= 1s^{-1}4mM - 1s^{-1}[S] = -1s^{-1}[S] + 4mMs^{-1}$
$y = ax + b$
$\Rightarrow$ Gerade von $(0, 4)$ zu $(4, 0)$

(c) Schnittpunkt von $v_D([S])$ mit Enzymkinetik ist $< 1mMs^{-1}$ bzw. ab $1mM$ wird Enzym inhibiert

(d) $[S] \approx 5mM$

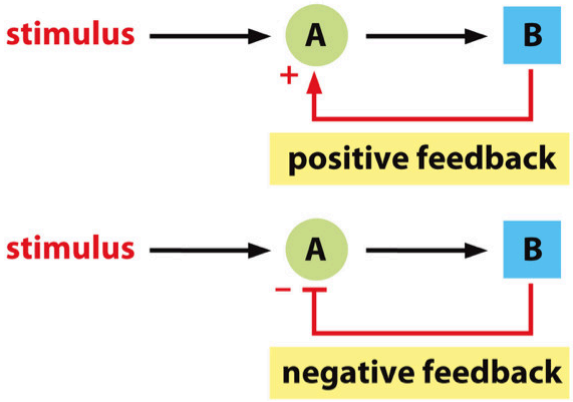
3 Modellierung von Stoffwechselwegen

3.1 Lernziele

Anhand von Beispielen Probleme aufzeigen, die in biologischen Pathways ohne Rückkopplungsmechanismen entstehen können und daraus die Notwendigkeit für solche Mechanismen erklären.

Der Input steigt in folgendem Pathway plötzlich um das sechsfache an: $\rightarrow A \xrightarrow{E_1}$. E_1 kommt nicht mit, $[A]$ steigt unbegrenzt an und führt zu einem *overshoot*. Flüsse in einem biologischen Pathway müssen regulierbar sein, um ungewollte Verhaltensweisen wie *overshoots* zu verhindern.

Erklären, welche Funktionen negative und positive Feedbacksysteme in Pathways innehaben.



Durch negative Feedbacksysteme stellt sich schnell ein bestimmter Gleichgewichtszustand ein. Durch positive Feedbacksysteme können kontinuierliche Inputsignalen in verschiedene diskrete Gleichgewichtszustände überführt werden.

An einem Beispiel erklären, wie ein System nur aufgrund eines dynamischen Prozesses eine Gedächtnisfunktion zeigen kann.

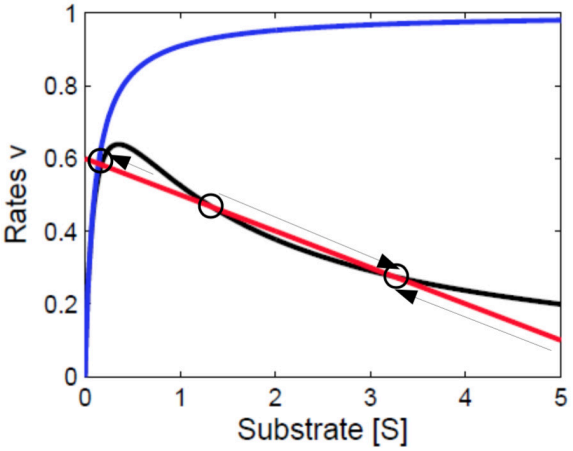


Abbildung 3.8: Mehrere Schnittpunkte zwischen der Geraden der Diffusion (rot) und der Kurve der Enzymkinetik (schwarz) bei bestimmten externen Substratkonzentrationen. Abhängig von der internen Substratkonzentration wird sich durch Feedback einer von zwei Schnittpunkten (äussere Kreise) einstellen.

Bei genau derselben externen Konzentration können sich zwei verschiedene *steady states* einstellen, wobei es von der aktuellen internen Substratkonzentation abhängt, welcher Gleichgewichtszustand erreicht wird. Der mittlere Schnittpunkt ist kein stabiler *steady state*.

Herausforderungen nennen, die in birektionalen biologischen Pathways (z.B. Glykolyse und Gluconeogenese) entstehen und die Notwendigkeit für Feedback-Mechanismen in solchen Pathways aufzeigen.

1) Damit die Richtung schnell gewechselt werden kann, müssen zu allen Zeiten bestimmte Mindestmengen von Enzymen und Cofaktoren vorhanden sein.

2) Sind gleichzeitig entgegengesetzte Reaktionen am laufen, verändert sich trotz hohen Energieaufwand wenig an der totalen Konzentration (*futile cycles*). Regulationsmechanismen verhindern dies.

Erklären, warum die möglichst gute Abschätzung der Parameterwerte beim Modellieren wichtig ist.

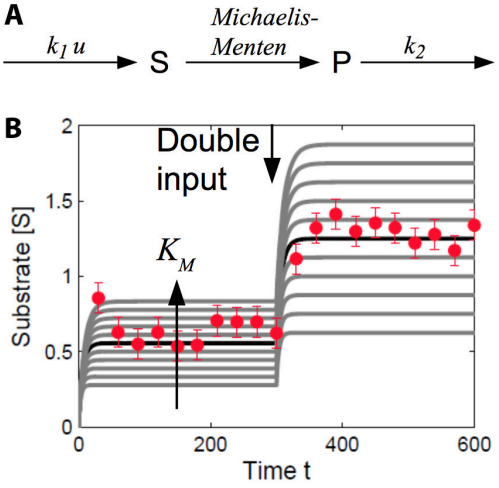


Abbildung 3.2: Abschätzung des K_M-Wertes durch Messung der Substratkonzentration. (A) Reaktionssytem, in dem die mittlere Reaktion enzymkatalysiert ist. **(B)** Da die Substratkonzentration abhängig vom K_M-Wert des Enzyms ist, kann man diesen durch Messung der Konzentration abschätzen.

Konzeptionell beschreiben, wie man die Qualität eines Parameterwertes bestimmt.
Man simuliert das System für verschiedene Parameterwerte und vergleicht die Resultate mit Messungen. Die Differenz sollte klein genug sein, dass sie bei angemessenem Signifikanzniveau auf Zufall zurückführbar ist.

Für eine gegebene Fragestellung zu den Kursthemen in Form eines kurzen Skripts in Pseudocode zeigen, wie man bei der Beantwortung vorgehen würde (kursübergreifendes Lernziel).
Universelle Schlüsselwörter wie IF, ELSE, DO oder FOR mit einfach verständlichem Englisch oder Deutsch kombinieren. Der genaue Syntax ist unwichtig, solange die Funktionsweise verständlich ist.

3.2 Beispielaufgaben

3.2.1

Im Stoffwechselweg $A \rightarrow B \rightarrow C$ gibt es eine Feedbackhemmung von C auf das erste Enzym, welches die Umwandlung von A zu B katalysiert. Nennen Sie 2 Gründe, wieso dieses Feedback für die Zelle notwendig sein könnte.

- kein Ressourcenverbrauch wenn genug C
- B oder C evtl. toxisch in grossen Mengen

3.2.2

Sie konstruieren einen synthetischen Stoffwechselweg zu einem biotechnologischen Produkt mit mehreren Enzymen, die sich alle durch Michaelis-Menten-Kinetik beschreiben lassen. Dieser Stoffwechselweg wird ohne Regulationsmechanismen in ein Bakterium eingebaut. (→ nächste Seite)

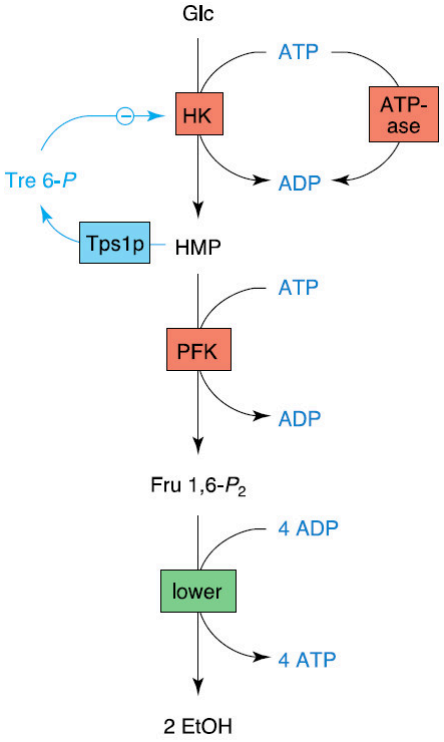
(a) Während des Produktionsprozesses verdoppelt sich der Zuckerfluss in das Bakterium schlagartig. Nennen Sie 2 mögliche Konsequenzen, die diese dynamische Veränderung für ihren synthetischen Stoffwechselweg haben könnte.

(b) Zu welchen konkreten Zellstoffwechselproblemen könnten diese Konsequenzen führen? Nennen Sie 3 Beispiele.

(c) Wie könnten Sie diese Probleme verhindern?

(a)
- ganzer Pathway schneller
- Überproduktion eines Intermediats

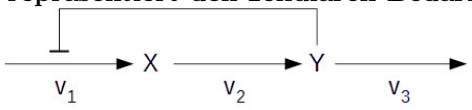
(b)
- futile cycles
- zu viel eines toxischen Intermediats
- substrate accelerated death



(c) Feedbackmechanismen

3.2.3

In einem Stoffwechselweg wie im Bild unten wird der Fluss v_1 durch das Endprodukt Y inhibiert; Fluss v_3 repräsentiert den zellulären Bedarf an Y.



(a) Welche Funktionen haben negative Feedbacksysteme im Allgemeinen?

(b) Welche Aussagen können Sie über die Konzentration von Y machen, wenn das System im stationären Zustand ist und Sie nicht mehr zu dem Stoffwechselweg wissen?

(c) Für ein vereinfachtes Modell des Stoffwechselwegs nehmen Sie an: $v_1 = \frac{k_1}{1+[Y]}$ und $v_2 = k_2[X]$, wobei $k_{1,2}$ kinetische Parameter und $[X]$, $[Y]$ Metabolitkonzentrationen sind. Geben Sie das ODE-Modell an.

(d) Mit dem Modell, dem gemessenen Fluss $v_3 = 1$ mmol/h und den Parameterwerten $k_1 = 2$ mmol/h und $k_2 = 1$ mmol/h, was sind die stationären Konzentrationen von X und Y?

(a)
- keine Ressourcenverschwendung
- keine übermässige Ansammlung eines Metaboliten

(b)
- $[Y]$ kann nicht hoch genug sein, um v_1 komplett zu inhibieren
- $[Y]$ bleibt konstant

(c) $\frac{d[X]}{dt} = v_1 - v_2 = \frac{k_1}{1+[Y]} - k_2[X] = 0$

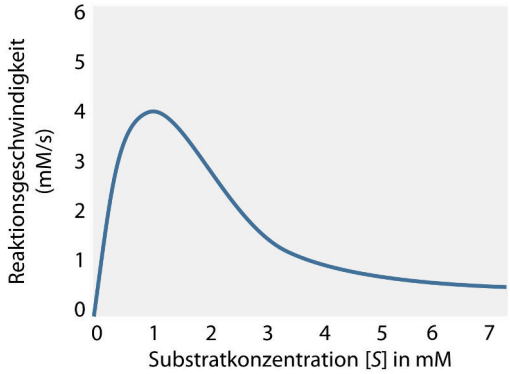
$\frac{d[Y]}{dt} = v_2 - v_3 = k_2[X] - v_3 = 0$

(d) $k_2[X] - v_3 = 0 \rightarrow k_2 = \frac{v_3}{[X]} \rightarrow [X] = \frac{v_3}{k_2} = 1$

$\frac{k_1}{1+[Y]} - k_2[X] = 0 \rightarrow [Y] = \frac{k_1}{k_2[X]} - 1 = 1$

3.2.4 (Fortsetzung von 2.2.3)

Ein Substrat S wird mit der Rate v_D von der Zelle über die Plasmamembran aufgenommen und in einem irreversiblen metabolischen Pathway weiterverwendet. Die Kinetik des ersten Enzyms des Pathways ist in der folgenden Abbildung gezeigt. Das System sei immer im *steady state*.



(a) Die Aufnahmegeschwindigkeit in die Zelle v_D sei gegeben durch die Gleichung $v_D = D([S_{\text{extern}}] - [S])$, wobei $D = 1s^{-1}$ eine Diffusionskonstante und $[S_{\text{extern}}]$ die Substratkonzentration ausserhalb der Zelle ist. Zeichnen Sie v_D als Funktion von $[S]$, wenn $S_{\text{extern}} = 4.5$ mM in das Diagramm ein.

(b) Welche interne(n) Substratkonzentration(en) sind möglich für $S_{\text{extern}} = 4.5$ mM? Begründen Sie Ihre Antwort und geben Sie eine ungefähre Abschätzung.

(c) Falls in (b) mehrere stationäre Zustände existieren, was ist der Mechanismus und wovon hängt ab, in welchem Zustand sich das System befindet?

(a) Gerade von $(0, 4.5)$ bis $(4.5, 0)$

(b) $[S]_1 \approx 1$; $[S]_2 \approx 3$

(c) je nach dem welchem Schnittpunkt $[S]$ näher ist

4 Regulation von Stoffwechselwegen

4.1 Lernziele

Eine biologische Erklärung für die Notwendigkeit von Regulationsmechanismen im Stoffwechsel geben. Wenn das Substrat plötzlich im übermass oder nicht mehr vorkommt stirbt die Zelle.

Verschiedene Mechanismen nennen, durch die Reaktionsraten von Enzymen reguliert werden, und jeweils angeben, welchen Faktor der Enzymkinetik sie beeinflussen. Genexpression, post-translationale Modifikation, allosterische Regulation, Enzym-Kapazität und das Metaboliten-Level

Angeben, auf welcher Zeitskala die verschiedenen Regulationsmechanismen stattfinden, und aufgrund von diesem Wissen erklären, wie man den Einfluss einzelner Regulationsprozesse untersuchen kann.
Man kann die Zeitskala trennen und Prozesse erst untersuchen, wenn sie beispielsweise schon im gehemmten *steady state* sind.

Erklären, wie die Metabolitkonzentration und metabolischer Fluss miteinander in Verbindung stehen.
Die Differenz aus Einfluss und Ausfluss bestimmt die Konzentrationsveränderung eines Metaboliten. Umgekehrt beeinflussen die Metabolitkonzentrationen die metabolischen Flüsse durch die Abhängigkeit der Reaktionsrate von der Substratkonzentration.

Verschiedene Feedback-Typen (proportional, integral, differentiell) beschreiben und ihre Funktion in der Regulation biologischer Pathways diskutieren.
Proportional: Durch allosterische Hemmung: PFK wird gehemmt, wenn die ATP Konzentration zu hoch wird. Jedoch ist die Hemmung zeitverzögert und das resultiert in Instabilität. Ein Problem ist, dass es nicht zum gleichen *steady state* führt. **Integral**: Korrigiert den Fehler als hätte es zu wenig ATP und viel AMP. Aktiviert PFK2 und die Glykolyse läuft weiter. **Differential**: Kompensiert für schnelle Veränderungen. Puffer sind differential. Z.B. die Keratin Kinase phosphoryliert Keratin, dephosphoryliert dieses aber wieder, wenn der ATP-Spiegel zu niedrig fällt.

4.2 Beispielfragen

4.2.1

Im Stoffwechselweg A → B → C gibt es eine Feedbackhemmung von C auf das erste Enzym, die Umwandlung von A zu B. Nennen Sie 3 Beispiele, wie diese Hemmung mechanistisch erreicht werden könnte und geben Sie eine kurze Erklärung bezüglich der zu erwartenden Geschwindigkeit.

- kompetitiv ($k_2[E]_0$, wie ohne Inhibition)
- nicht kompetitiv (allosterische Hemmung) ($\frac{v_{max}}{1+\frac{[I]}{K_I}}$)
- unkompetitiv ($\frac{v_{max}}{1+\frac{[I]}{K_I}}$)

4.2.2

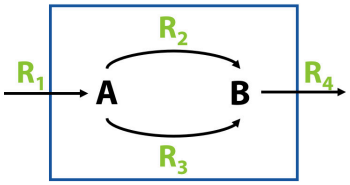
Die irreversible Reaktion S → P wird vom Enzym E mit dem katalytischen Koeffizienten k_{cat} und der Substrataffinität k_m katalysiert. Die Reaktion folgt einer Michaelis-Menten-Kinetik und der Fluss wird durch die folgende Gleichung beschrieben: $v = [E] * k_{cat} * \frac{[S]}{k_m+[S]}$. Nehmen Sie an, dass System wäre transkriptionell reguliert. Welcher Teil der Gleichung wird sich ändern und wieso? Wird der Fluss v grösser, kleiner oder bleibt er unverändert, wenn sich das Expressionlevel auf 10% des Normalzustandes reduziert (unter der Annahme, dass von jedem Transkript ein Protein gebildet wird)?
Fluss kleiner, da $[E]$ transkriptionell Reguliert wird.

5 Analyse von metabolischen Netzwerken durch Flux Balance Analysis

5.1 Lernziele

Die grundlegende Annahme kennen, die man bei der Modellierung sehr grosser Stoffwechselnetzwerke trifft und können erklären, warum diese Annahme notwendig ist.
Man nimmt an, dass alle Metabolitkonzentrationen im *steady state* sind. Man kennt die Stöchiometrie aller Reaktionen und sucht nach den Werten der metabolischen Flüsse unter bestimmten Bedingungen. Für ein dynamisches Modell müsste man alle Informationen über jede zelluläre Reaktion eines Organimus haben. Ausserdem ist die benötigte Rechenleistung heutzutage noch zu gross.

Konzeptionell beschreiben, wie man von einzelnen biochemischen Reaktionen genomweite Stoffwechselnetzwerke rekonstruiert und welche Quellen man dafür nutzt.



$$\frac{d[A]}{dt} = v_1 - v_2 - v_3 = 0 \qquad v_1 = v_2 + v_3$$
$$\frac{d[B]}{dt} = v_2 + v_3 - v_4 = 0 \qquad v_4 = v_2 + v_3$$

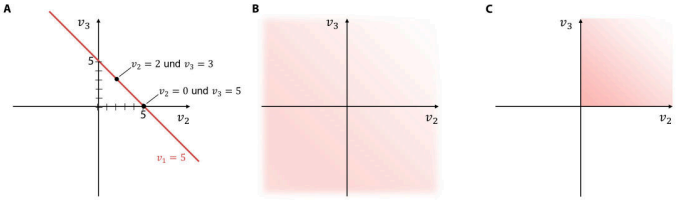
Die stöchiometrische Matrix für das Schema eines einfachen metabolischen Netzwerks aufstellen und angeben, in welchem Zahlenverhältnis die Flüsse im Netzwerk zueinander stehen.

$$\boldsymbol{S} = \begin{matrix} R_1 & R_2 & R_3 & R_4 \\ \downarrow & \downarrow & \downarrow & \downarrow \end{matrix} \begin{pmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \end{matrix}$$

Pro Zyklus entsteht A ein Mal in R_1 und wird in R_2 ein Mal verbraucht usw.

$$\frac{dc}{dt} = S * v = 0 \quad \text{mit} \quad v = (v_1, v_2, v_3, v_4)$$

Den zweidimensionalen Lösungsraum für zwei Flüsse aus einem einfachen metabolischen Netzwerks grafisch darstellen.



(A) Für jeden Einfluss (hier $v_1 = 5$) liegen alle möglichen Lösungen für v_2 und v_3 auf einer Geraden im Lösungsraum.
(B) Weil v_1 beliebig gross sein kann, setzt sich der Lösungsraum für v_2 und v_3 aus unendlich vielen Parallelen Geraden zusammen und ist eine unendlich grosse Fläche.
(C) Unter der Annahme, dass irreversible Flüsse nicht kleiner als 0 sein können, lässt sich der Lösungsraum auf den ersten Quadranten einschränken.

Eine Kernel-Matrix für ein einfaches metabolisches Netzwerk konstruieren und zur Analyse von Kopplungen in dem Netzwerk verwenden. Den Zusammenhang zwischen dem Lösungsraum einer stöchiometrischen Matrix und ihrer Kernel-Matrix erklären.

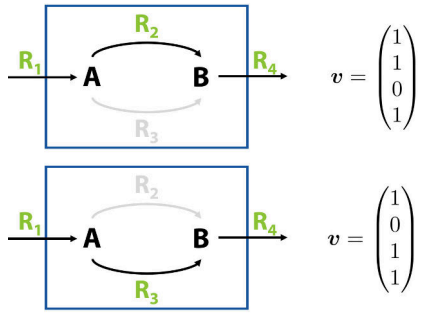


Abbildung 5.5: Zwei Lösungen für die Flussvektoren unseres Reaktionssystems.

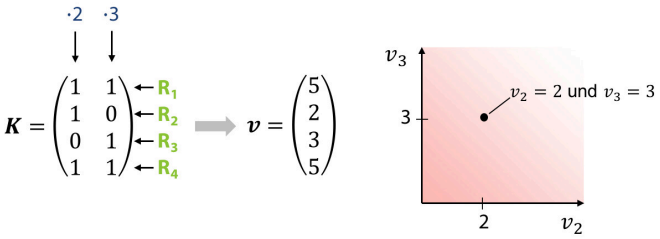


Abbildung 5.6: Zusammenhang zwischen der Kernel-Matrix eines Gleichungssystems und dem Lösungsraum. Jede Spalte der Kernel-Matrix ist eine linear unabhängige Lösung des Gleichungssystems. In unserem Reaktionssystem gibt es zwei solche Lösungen für die vier Flüsse der Reaktionen R_1 bis R_4. Durch Linearkombination der Spalten der Kernel-Matrix kann man jeden beliebigen Punkt im Lösungsraum beschreiben. Hier ist der Lösungsraum der beiden Flüsse v_2 und v_3 gezeigt und für einen Punkt, durch welche Linearkombination der Vektoren man diesen erreicht.

Prinzipienbasierte und datenbasierte Einschränkungen (*constraints*) nennen, die bei einer Flux Balance Analysis getroffen werden können.

1) Struktur des Netzwerks
2) *steady-state*-Annahme
3) (Ir)reversibilität der Reaktionen
4) Obergrenzen der Reaktionsraten
5) Bedingungen unter denen sich die Zelle befindet.

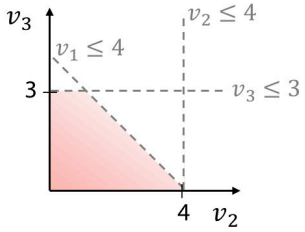


Abbildung 5.7: Einschränken des Lösungsraums durch Setzen von Obergrenzen.

Die Notwendigkeit für Zielfunktionen (*objective functions*) bei einer Flux Balance Analysis erläutern und die Eignung verschiedener Zielsetzungen für eine gegebene Fragestellung diskutieren.

Allgemein ausgedrückt gibt die Zielfunktion einen Wert dafür, wie gut ein Stoffwechselnetzwerk darin ist, ein bestimmtes Ziel zu erreichen (z.B. Sekretion eines bestimmten Stoffes, maximale Wachstumsrate, Robustheit bei wechselnden Umweltbedingungen, maximale Effizienz bei der Nutzung einer Ressource). Oft optimiert man statt nur einen mehrere Flüsse gleichzeitig.

Allgemein beschreiben, was Input und Output einer Flux Balance Analysis sind.
Nährstoffe sind der Input und Biomasse ist der Output.

An Anwendungsbeispielen von Flux-Balance-Analysen aufzeigen, welche Erkenntnisse jeweils durch die Analyse gewonnen wurden.

- Vorhersage von Mutantenverhalten: Stöchiometrisches Modell für *E. coli*; FBA für maximales Wachstum nach Gendeletion; 86% Vorhersagewahrscheinlichkeit
  Annahmen: Zellen wollen so schnell wie möglich wachsen und sind an dieses Ziel angepasst.
  FBA-Vorhersagen sind mit Experimenten einig, aber nicht direkt nach einem Knock-out.

Anhand von Beispielen begründen, wie die Formulierung einer biologischen Fragestellung die Komplexität der computergestützten Analysen beeinflussen kann und ggf. eine solche Analyse verhindern kann.
FBA ist nicht möglich, wenn folgende Situationen auftreten:

- parallele Pathways
- *futile cycles*
- reversible Reaktionen
- verzweigte Pathways ⇒
  - Einschränkungen einführen
  - stark vereinfachende Annahmen treffen

Anhand von Beispielen diskutieren, wie experimentelle Daten für die Analyse von Stoffwechselmodellen integriert werden können und welche (prinzipiellen) Schwierigkeiten dabei auftreten.

- **Omics-Daten**: Enzymmenge in Enzymkinetiken sind linear → $\frac{dc(t)}{dt} = S * E * v' = 0$
  Schwierigkeiten
  - $[E]$ verändert sich aber der Fluss nicht (Metabolitenkonzentration passt sich an)
  - Einschränkungen erlauben keinen metabolischen *steady state*
- **Regulatorische Netzwerke**: Reaktionen sind von transkriptionellen Netzwerken reguliert → einfache Zustände (an/aus ⇒ 1/0) als Einschränkungen einführen

Für eine gegebene Fragestellung zu den Kursthemen in Form eines kurzen Skripts in Pseudocode zeigen, wie man bei der Beantwortung vorgehen würde (kursübergreifendes Lernziel).

Konzept Flux Variability Analysis (FVA): Minimalen und maximalen Fluss von Reaktionen in einem Netzwerk finden, während ein bestimmter Zustand beibehalten wird.

Inputs	Stöchiometrische Matrix als auch Ober- und Untergrenzen der Flüsse
Output	Minimale und Maximale Flüsse

```
FOR each of the reaction directions {forward,
    backward}
  FOR each reaction as a target
    SET weights for FBA such that the target reaction
        is the objective in the correct direction
    RUN FBA
    SAVE optimized flux for the target reaction
  END
END
```

5.2 Beispielfragen

5.2.1

Sie haben ein stöchiometrisches Netzwerkmodell für den Menschen mit m Metaboliten und n enzymkatalysierten Reaktionen gegeben. Sie wollen FBA anwenden, um metabolische Funktionen des Netzwerkes zu analysieren.

(a) Geben Sie ein kleines aber effizientes Skript in Pseudocode an, das es Ihnen erlaubt, herauszufinden welche Doppelmutanten letal für Zellwachstum sind. Berücksichtigen Sie dabei alle letalen Doppelmutanten, nicht nur synthetisch letale.

(b) Wie müssten Sie Ihr Skript anpassen, wenn Sie nur an synthetisch letalen Doppelmutanten interessiert sind? Antworten Sie qualitativ, nicht durch Angabe eines neuen Skripts.

(c) Welche Objective Function würden Sie nicht wählen, um Vorhersagen über Flussverteilungen zu erhalten (und warum)?

(d) Aus Genexpressionsanalysen wissen Sie, dass bestimmte Enzyme in bestimmten Geweben nicht exprimiert werden. Wie könnten Sie dieses Wissen nutzen, um das Modell zu verfeinern?

(a) **Inputs**: Stöchiometrisches Modell des Stoffwechselnetzwerks, Einschränkungen der Reaktionsraten, Optimierungsfunktion
Output: Kombinationen von je 2 Reaktionen die letal sind, wenn beide Reaktionen fehlen

```
Create a matrix of zeros M with the size n x n
  For each (but the last) reaction k
    For each of the subsequent reactions l
      Load model
      Set upper and lower flux bounds for
          reactions k and l to zero
      Run FBA
      If combination is lethal
        Set value k, l in matrix M to 1
    End
  End
```

(b) vorher FBAs mit je nur einer Mutante laufen lassen und lethale später nicht mehr berücksichtigen

(c) evolutionär sinnvoll für Organismus, da lethale Doppelmutanten keinen evolutionären Sinn ergeben

(d) die von diesen Enzymen katalysierten Reaktionen von Anfang an ignorieren

5.2.2

Sie analysieren ein metabolisches Netzwerk im stationären Zustand. Das Netzwerk hat fünf interne Metabolite $A-E$, vier interne Reaktionen mit Flüssen $r_1 - r_4$, sowie zwei externe Reaktionen mit Flüssen r_A und r_E. Alle Reaktionen sind irreversibel; sie lauten:

$$\xrightarrow{r_A} A \qquad A \xrightarrow{r_3} 2D$$
$$A \xrightarrow{r_1} B \qquad C + D \xrightarrow{r_4} A + E$$
$$B \xrightarrow{r_2} C \qquad e \xrightarrow{r_E}$$

(a) Geben sie die stöchiometrische Matrix des Netzwerkes an.

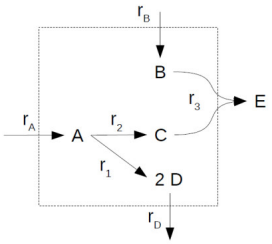
(b) Sie haben den Fluss $r_E = 1mmol/h$ gemessen. Welche anderen stationären Flüsse können Sie direkt bestimmen und welche Werte haben diese?

(a)
$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} r_A & r_1 & r_2 & r_3 & r_4 & r_E \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \end{array}$$

(b)
$$r_E = 1mmol/h \quad \rightarrow \quad r_A = 1mmol/h$$
$$r_4 = 2mmol/h$$
$$r_1 + r_3 - (r_4 - 1mmol/h) = 1mmol/h$$
$$r_1 + r_3 = 2mmol/h$$
$$r_1 = r_2$$

5.2.3

Sie analysieren das gegebene metabolische Netzwerk im stationären Zustand. Pfeile notieren Reaktionen (alle Reaktionen sind irreversibel, die entsprechenden Symbole für Flüsse sind annotiert) und Buchstaben die Metabolite. Die Zellgrenze ist durch die gestrichelte Linie gezeigt.



(a) Geben Sie die stchiometrische Matrix des Netzwerkes an.

(b) Sie haben die Aufnahmeraten $r_A = 2mmol/h$ und $r_B = 1mmol/h$ gemessen. Welche anderen stationären Flüsse können Sie bestimmen und welche Werte haben diese?

(a)
$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} r_A & r_B & r_1 & r_2 & r_3 & r_D \end{array} \\ \begin{pmatrix} 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 2 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

(b) $r_A = 2mmol/h; \quad r_B = 1mmol/h$
$r_1 + r_2 = 2mmol/h$
$r_3 + r_D = 3mmol/h$
$r_D = 3 * r_1$

5.2.4

Sie haben ein stöchiometrisches Netzwerkmodell für humane Krebszellen mit Metaboliten und Enzym-katalysierten Reaktionen gegeben. Sie wollen FBA anwenden, um metabolische Funktionen des Netzwerkes zu analysieren.

(a) Mit diesem Modell ist kein Zellwachstum möglich. Sie vermuten, dass eine Komponente der im Modell repräsentierten Biomasse nicht synthetisiert werden kann. Geben sie ein kleines Skript in Pseudocode an, das es Ihnen erlaubt, herauszufinden welche Biomassekomponente dies ist.

(b) Nach Korrektur des Modells: welche Objective Function würden Sie wählen, um Vorhersagen über Flussverteilungen zu erhalten (und warum)?

(c) Wie könnten Sie das Modell verwenden um zu analysieren, welche metabolischen Reaktionen für diese Krebszellen essentiell sind?

(a) **Input**: stöch Mod. des Stoffwechselnetzwerkes, Einschränkungen der Reaktionsraten, Optimierungsfunktion
Output: Vektor der für jede Reaktion, die im Modell 0 ist, besagt, ob sie beim Anschalten Biomassensynthese erlaubt

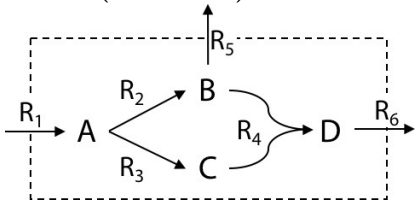
```
FOR each reaction that is 0 in the model
  SET reaction to 1 (running)
  LOAD model
  RUN FBA
  IF biomass is synthesized
    SAVE reaction as crucial in vector
END
```

(b) maximales Wachstum

(c) je eine Reaktion löschen und testen, ob Zelle noch wächst

5.2.5

Gegeben ist das folgende Reaktionssystem, das aus vier Metaboliten (A, B, C, D) und sechs irreversiblen Reaktionen ($R1$ bis $R6$) besteht:

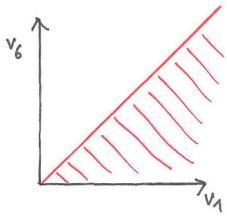


(a) Stellen Sie die Massenbilanz für jeden der vier Metaboliten auf.

(b) Zeichnen Sie den Lösungsraum für die Geschwindigkeit v_S in Abhängigkeit von v_4 in das Diagramm.

(a) $\frac{d[A]}{dt} = v_1 - v_2 - v_3 \qquad \frac{d[B]}{dt} = v_2 - v_4 - v_5$
$\frac{d[C]}{dt} = v_2 - v_4 \qquad \frac{d[D]}{dt} = v_4 - v_6$

(b)



5.2.6

Was ist die Hauptannahme bei einer Flux Balance Analysis? Erklären Sie anhand dieser Annahme den Namen der Analysemethode.
alles im *steady state*; Flux balance = Fluss-Gleichgewicht → keine Veränderungen der Flüsse

7 Modellierung von Signaltransduktionswegen

7.1 Lernziele

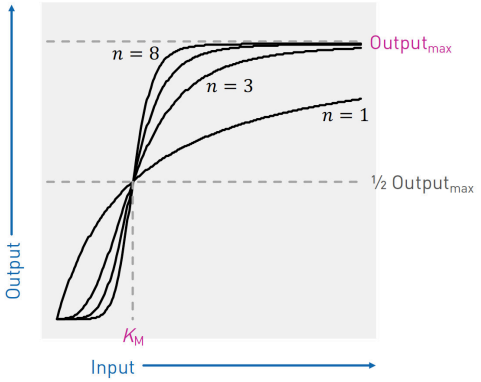
Herausforderungen nennen, die für eine Zelle bei der Signaltransduktion auftreten. Sie muss auf sehr kleine Konzentrations-Änderungen reagieren können. Z.B. bei der Michaelis-Menten-Kinetik muss der Input um das 81-fache gesteigert werden um von <10% auf mehr als >90% zu steigen. Da Zellen schneller als das reagieren müssen, gibt es Ultrasensitivität.

Ultrasensitivität definieren und verschiedene Mechanismen beschreiben, durch die eine Zelle Ultrasensitivität erreicht. Ultrasensitivität bedeutet, dass ein System stärker auf eine Signalveränderung reagiert als durch Michaelis-Menten Kinetik. So kann schalterartiges Verhalten erzielt werden. Mechanismen dafür sind

- Kooperation (Bsp. Hämoglobin)
- Signalkaskaden (MapKKK)

Eine biologische Interpretation verschiedener Hill-Koeffizienten in der Hill-Gleichung geben und jeweils den Kurvenverlauf zeichnen.

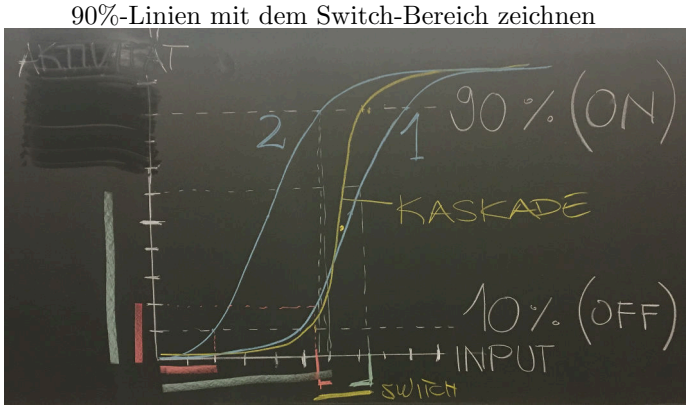
$$\text{Output} = \text{Output}_{\max} \frac{\text{Input}^n}{\text{Input}^n + K_M^n}$$



Wenn $n = 1$ liegt Michaelis-Menten-Kinetik vor. n ist der sogenannte Hill-Koeffizient, ist ein Mass für Ultrasensitivität und beschreibt die kooperative Bindung. Je grösser n, desto steiler ist die sigmoidale Kurve und desto kleiner ist der Änderungsbereich beim Inputwert, der zu einer schalterartigen Veränderung des Outputs führt. $\text{Input}_{0.5}$ ist der Inputwert, bei dem die Hälfte des maximalen Outputs erreicht wird.

Grafisch zeigen, warum die Kombination mehrerer ultrasensitiver Schritte in einem Pathway zu einem noch höheren Hill-Koeffizienten der gesamten Kaskade führt. n_{tot} ist grösser als n_1 und n_2 zusammen. K_{M1} bestimmt $K_{M,tot}$ der Kaskade (kann nicht kleiner sein als das K_M der ersten Reaktion), denn die Schalterreaktion liegt im Bereich des ersten Proteins.

1) Inputwert (x-Achse) des Schnittpunktes von 10% mit Kurve 2 ablesen
2) Inputwert auf Aktivitätsachse (y) übertragen und Schnittpunkt mit Kurve 1 suchen
3) Schnittpunkt auf x-Achse übertragen
4) Schritte 1) bis 3) für 90% wiederholen
5) die letzten 10%- bzw. 90%-Schnittpunkte bilden auf der x-Achse den Switch-Bereich
6) Sigmoidale Kurve durch Schnittpunkte der 10%- und 90%-Linien mit dem Switch-Bereich zeichnen



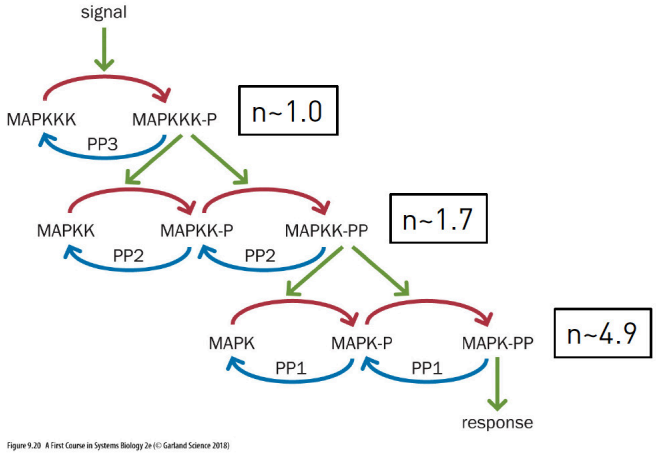
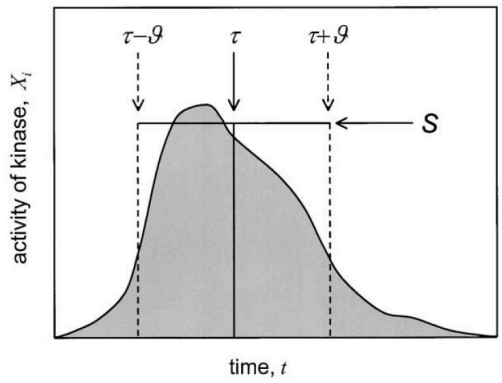


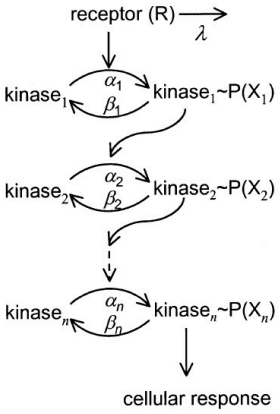
Figure 9.20 A First Course in Systems Biology 2e (© Garland Science 2018)

Kriterien nennen, durch die man das Ergebnis einer Signalkaskade quantitativ charakterisieren kann.

- Signaling time τ: Expected time of arrival of the signal.
- Signal duration υ: Variance of the expected time.
- Signal amplitude S: Average amplitude during signaling.



Diskutieren, welche Faktoren die Dynamik einer Signalkaskade beeinflussen.



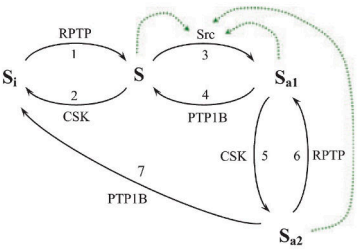
$$\frac{dX_1}{dt} = v_{p,1} - v_{d,1}$$
$$= \alpha_1 \widetilde{X}_1 R - \beta_1 X_1$$
$$\frac{dX_i}{dt} = v_{p,i} - v_{d,i}$$
$$= \alpha_i \widetilde{X}_i X_{i-1} - \beta_i X_i$$

X: phosphorylierte Kinase
$\widetilde{X}$: nicht phosphorylierte Kinase

Erklären, warum eukaryotische Signalkaskaden aus mehreren Kaskadestufen bestehen. So kann das Signal amplifiziert oder abgeschwächt werden. Es wird auch enorm verschnellert.

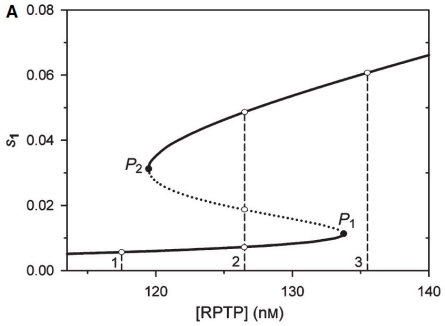
Am Beispiel eines Signalmoleküls erklären, wie ein Signalmolekül viele verschiedene Signale interpretieren und durch unterschiedliche dynamische Verhaltensweisen reagieren kann. Src-Kinase: Schlüssel Regulator mit mehreren Funktionen: Beweglichkeit, Proliferation, überleben, Zell-Zell Adhäsion. Bis zu 350 potenziellen Substraten.

- S_i: inhibited (pY527, Y416)
- S: partial act. (Y527, Y416)
- S_{a1}: active (Y527, pY416)
- S_{a2}: active (pY527, pY416)



N. Kaimachnikow & B. Kholodenko (2009) *FEBS J.* 276: 4102.

Wird reguliert durch 2 Phosphorilierungs-Events, von dene eines inhibiert und das andere aktiviert. So gibt es 4 Stadien und insgesamt 7 Reaktionen. Sie folgen nicht der Michaelis-Menten-Kinetik. So ergibt sich ein dynamisches Gedächtnis, welches vom vorherigen Zustand der Kinase abhängt. **RPTP** ist die Phosphatase, welche die inaktive Form der Src-Kinase Si durch Dephosphorylierung in die teilweise aktive Form S überführt.



Ist die Konzentration von RPTP tief, ist auch die Konzentration der Form S und damit die Src-Aktivität tief (Punkt 1). Ist die Konzentration von RPTP hingegen hoch, liegt viel der aktiven Form der Src-Kinase vor, die sich durch Autophosphorylierung selbst weiter aktivieren kann (Punkt 3).

Interessant wird es, wenn man die Src-Kinase-Aktivität bei einer mittleren RPTP-Konzentration betrachtet: Abhängig von der vorherigen Aktivität der Src-Kinase ist die Aktivität nun tief, wenn sie auch vorher tief war, bzw. hoch, wenn sie vorher hoch war (Punkt 2). Die Schnittpunkte sind dabei *steady states*.

$$v_3 = \left(\frac{k_S^{cat}}{K_S}[S] + \frac{k_{a1}^{cat}}{K_{a2}}[S] + \frac{k_{a2}^{cat}}{K_{a2}}[S] \right)[S]$$

Erklären, was bei einer Sensitivitätsanalyse untersucht wird und an einem Beispiel erklären, wie man durch so eine Analyse potentielle Angriffspunkte für Medikamente identifizieren kann. Man verändert die Parameter K_M für die verschiedenen Reaktionsschritte und sieht anhand des Outputs wie sensitiv dieser Teilschritt ist. Beim sensitivsten Teilschritt findet die grösste Veränderung des Outputs statt. Dies zeigt den Angriffspunkt des Medikaments an. Krebs entsteht häufig durch eine Mutation im MapKKK Pathway oder im AKT Pathway. Man sucht den sensitivsten Teilschritt/Rezeptor/Kinase und versucht ihn mit einem Medikament zu inhibieren.

7.2 Beispielfragen

7.2.1

Die Aktivität einer Signalkaskade im stationären Zustand kann mit Hilfe der sogenannten Hill-Kinetik beschrieben werden:

$$A_i = \frac{A_{i-1}^n}{K_i^n + A_{i-1}^n}$$

wobei A_M die Aktivität der i-ten Kinase ist, welche von der vorhergehenden Kinaseaktivität A_{i-1} abhängt, n der Hill-Koeffizient (Exponent) ist und K_M die Michaelis-Menten-Konstante für die Aktivierung von Kinase i ist.

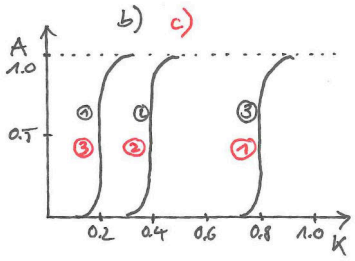
(a) Wie muss n sein, um Ultrasensitivität jedes Signalschrittes zu erreichen?

(b) Nehmen Sie an, dass eine MAPK-Kaskade mit $i = 1 \dots 3$ Kinasen die folgenden Parameterwerte hat: $n = 8$, $K_1 = 0.2$, $K_2 = 0.4$ und $K_3 = 0.8$. Skizzieren Sie graphisch die Aktivitäten der einzelnen Kinasen als Funktion der vorhergehenden Kinaseaktivität bzw. des Inputs A_0 (alle Signale sind in $[0, 1]$).

(c) Welche Kinasekaskade wird den Output auf der letzten Ebene bei einem geringeren Inputsignal aktivieren: die Kaskade aus (b) oder ein alternativer Signalweg mit $n = 8$, $K_1 = 0.8$, $K_2 = 0.4$ und $K_3 = 0.2$? Begründen Sie Ihre Antwort.

(a) > 1

(b)



(c) (b) ist schneller, da bei (c) zuerst 0.8 erreicht werden muss, bevor Kinasen 2 und 3 aktiv werden

8 Analysis of omics data

Know what biological insights are to be obtained by data mining of omics data. Man kann die Daten ordnen und findet evt. etwas. *Data mining* wird auch benutzt um Hypothesen zu generieren.

Describe the typical problems associated with omics data. Es sind sehr viele Daten und es ist schwierig den Überblick nicht zu verlieren (Unwichtiges rausfiltern). Das Risiko des *overfitting* besteht. Die Experimente müssen technisch reproduzierbar bleiben.

$$\#\text{ of detected features} >>> \#\text{ of samples}$$

Describe the role of statistics in biological context. Statistik wird benutzt, um Hypothesen zu verwerfen. Es kann jedoch nie mit 100% Sicherheit etwas bewiesen oder verworfen werden.

8.0.1 Data Mining

1) biologische Frage
2) keine eindeutige Antwort erwarten
3) positive und negative Kontrolle
4) möglichst einfach halten
5) verschiedene Herangehensweisen nutzen

8.0.2 Analyse und Statistik

Differential, univariate analysis between two groups Die Daten wurden bereits in zwei Gruppen z.B. Mutant/WT oder gesund/krank gesammelt. Nun ist interessant, ob weitere Abweichungen wie zelluläre Prozesse bestehen.

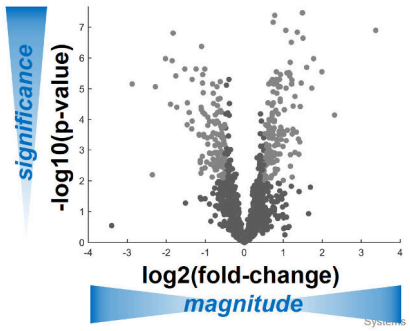
Understand the basic form of univariate analysis. Man nimmt eine Komponente zu einer bestimmten Zeit und vergleicht die zwei Gruppen. Dies wird für jede Komponente widerholt. Man berechnet die Grösse der Veränderung mit den *fold-change*. Für bessere symmetrische Visualisierung benutzt man den Logarithmus davon.

$$FC = \frac{mean(GroupA)}{mean(GroupB)}$$
$$log_2(FC) = log_2\left(\frac{mean(GroupA)}{mean(GroupB)}\right)$$

Man benutzt dann verschiedene statistische Test um herauszufinden, ob die Null-Hypothese (das die zwei Gruppen derselben Verteilung folgen) stimmt.

- **t-Test**: beruht auf Normalverteilug
- **Mann-Whitney- / Wilcoxon- / Ranksum-Test**: benutzen Ranking
- **Permutations-Test**: Für Studien mit vielen Proben; Die Datenpunkte werden zufällig einer Gruppe zugewiesen und es wird getestet, wie oft die neuen Gruppen mehr Sinn machen als die ursprünglichen.

Mit dem Generierten p-Wert wird dann ein *volcano plot* erstellt. Allgemein gilt in der Biologie: $p < 0.05$ ist akzeptierbar; $p < 0.01$ ist super.

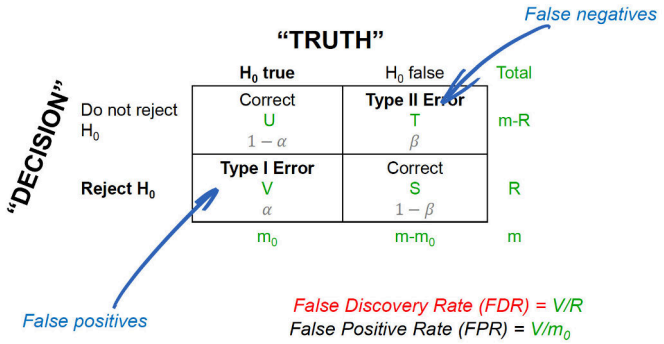


Fold Change Treshold

- beliebiges Signifikanzniveau bestimmen (z.B. $|log_2(FC)| > 1$)
- empirisches Vorgehen: Niveau so wählen, dass ähnliche Dinge nicht signifikant unterschiedlich sind
- Schätze die Verteilung zwischen beliebigen oder allen Teilmengen der negativen Kontrollen

Multiples Testen: Das Problem mit dem p-Treshold ist, dass mit einem Signifikanzniveau von 0.05 eine Chance von 5% auf ein falsch positives resultat besteht. An sich ist das in Ordnung bei einem durchgeführten Test, wenn man den Test jedoch oft widerholt, steigt die Wahrscheinlichkeit auf einen falsch positiven Wert rasant an. Es gibt drei Möglichkeiten dies zu Korrigieren.

- **Family-Wise Error Rate**: Schützt gegen alle falschen Positive und ist somit sehr strikt. Bonferroni: Korrigiere den p-Wert mit der Anzahl Tests $p \leq \frac{\alpha}{m}$; Wahrscheinlichkeit min. 1 falsches Positiv zu erhalten $= FWER = P(V \geq 1)$
- **False Discovery Rate (FDR)**: Steuert die Anzahl falsch positiver Ereignisse unter den abgelehnten Hypothesen. p-Wert > Korrigierter p-Wert der FDR, verwirft H_i wenn korrigiertes $p_i \leq \alpha$; OK wenn Differenzen selten ($pi0 < 1$).
- **Positive False Discovery Rate**: Kontrolliert die Rate der falschen Ereignisse. Besser wenn Differenzen oft vorkommen ($pi0 < 1$)

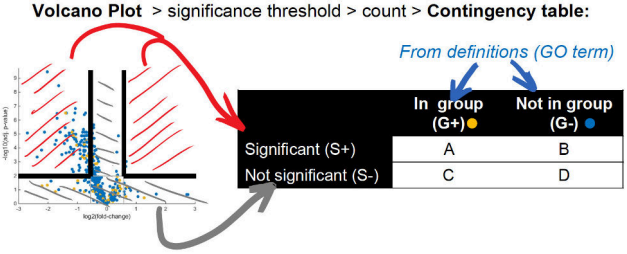


0 signifikante Veränderungen	Datenqualität prüfen; niedrige p-Werte, hoher FC → Replikatanzahl erhöhen
wenig signifikante Veränderungen	einfachster Fall → was ist die Identität der Marker?
viele signifikante Veränderungen; kein klares Ranking	relativ häufig; schwierig Hypothese aufzustellen; Kombination aus primären und sekundären Effekten → *enrichment analysis*

8.0.3 Enrichment Analysis

Wie erkennt man signifikant veränderte zelluläre Prozesse? Ordne die gemessenen Eigenschaften in Untergruppen. Eigenschaften können dabei in mehreren Gruppen gleichzeitig vorkommen. Sind bestimmte signifikante Eigenschaften in einer gewissen Gruppe besonders häufig? → *hotspot*

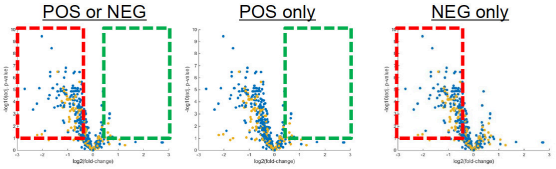
Volcano Plot > significance threshold > count > **Contingency table:**



Damit kann man nun einen Fischer Exact Test machen und herausfinden, wie gross die Wahrscheinlichkeit ist die gleichen Resultate unter der Nullhypothese zu erhalten.

8.0.4 Gen Set Enrichment Analysis (GSEA)

1) entscheiden welche Anzeichen miteingeschlossen werden



2) Vorgehen für 1 Pathway / Gruppe / Prozess
  1) sehr lockere Grenzen setzen (z.B. $|log_2(FC) > 0.2$ und $p < 0.1$)
  2) alle Ergebniss nach p-Wert (bevorzugt) oder $log_2(FC)$ ordnen
  3) *contingency tables* für 2, 3, 4, ... alle besten Ergebnisse erstellen
  4) jeweils p-Wert mit *Fisher's exact test* berechnen
  5) niedrigsten p-Wert behalten = bestes *enrichment*
3) für alle Pathways / Gruppen / Prozesse wiederholen
4) FDR-Korrektur (Benjamini-Hochberg oder Storey)

9 Feature Selection

Das grundlegende Problem der Feature Selection in der Systembiologie ist es, diejenigen Komponenten eines Systems zu finden, die einen bestimmten Output, Funktion oder Phänotyp beeinflussen. Gründe dafür sind um melukulare Mechanismen des Phänotyps zu verstehen, die Dimension des Modells zu verringern und um allfälliges Rauschen zu entfernen. Ein Anwendungsbeispiel sind GWAS (siehe Vorlesung Genetik, Genomik, Bioinformatik für Einzelheiten).

9.1 Univariate Feature Selection

Für jedes Feature j wird der Relevanz-Wert (*score*) $r(j)$ berechnet. Anschliessend werden Features nach $r(j)$ sortieret und als geordnete Liste zurückgeben. Limitationen sind:

- nur Effekt eins Gens
- ignoriert Korrelation zwischen Genen
- ignoriert additive Effekte zwischen Genen
- ignoriert Interaktionen zwischen Genen

Univariate Feature Selection ignoriert den systembiologischen Charakter des Problems.

9.1.1 Anzahl Features

- Ein zufälliges Rausch-Feature z generieren. Alle Features mit $r(j) > r(z)$ wählen.
- Für jeden Zusammenhang zwischen Feature und Phänotyp einen p-Wert berechnen. Nur Features mit signifikanten Zusammenhängen wählen.

9.1.2 Multiples Testen

Durch Zufall werden bei Tausenden getesteten Feature einige mit fälschlicherweise signifikanten Zusammenhängen dabei sein. Die Bonferroni-Korrektur kann diese Problem lösen, ist oft aber zu streng. Alternativen sind z.B. die *False Discovery Rate*.

9.1.3 Instabilität

Instabilität von Resultaten: Es werden verschieden Teilmenge eines Datensatzen getestet oder unterschiedliche Feature-Selection-Methoden verwendet. Die resultierenden Rangordnungen können stark unterschiedlich voneinander sein. Um aus mehreren Resultaten ein möglichst gutes zu gewinnen, können folgende Strategien angewandt werden:

- Durchschnittlichen Rang eines Gens über alle Experimente hinweg berechnen.
- Die Wahrscheinlichkeit berechnen, dass ein Gen in jedem Experiment unter den Top-k ist (z.B. mit **Fisher's Inverse χ^2 test**)

9.2 Multivariate Feature Selection

9.2.1 Additive Models

Methoden mit linearer Regression, die aus x eine Vorhersage für y zu treffen versuchen. Features bekommen Gewichtungen in β. Relevante Features haben eine Gewichtung, die ungleich null ist. Je nach Formulierung des Modell, trifft dies aber auf sehr viele Features zu.

$$arg\ min_\beta ||y - X\beta||_2^2$$

9.2.2 L1- und L2-Normen

Die L2-Norm belohnt höhere Werte statt kleinere zu verringern. Die L1-Norm belohnt beides gleich stark.

9.2.3 Lasso Model

Lösungen mit wenig relevanten Features werden bevorzugt. Bei Gruppen von korrelierten Features wird aber nur eines davon ausgewählt.

$$arg\ min_\beta ||y - X\beta||_2^2 + \gamma_1 ||\beta||_1$$

Die L1-Norm von β wird minimiert: $||\beta||_1 = \sum_{i=1}^{d} |\beta_i|$

9.2.4 Ridge Regression

Lösungen in denen korrelierte Features ähnliche Gewichtungen erhalten, werden bevorzugt. Die Lösung ist oft aber nicht besonders spärlich.

$$arg\ min_\beta ||y - X\beta||_2^2 + \gamma_2 ||\beta||_2^2$$

Die L2-Norm von β wird minimiert: $||\beta||_2 = \sqrt{\sum_{i=1}^{d} \beta_i^2}$

9.2.5 Elastic Net

Lösungen in denen Gruppen von korrelierten Features ähnliche Gewichtungen haben und wenige Gewichtungen ungleich null sind. Es müssen allerdings zwei Parameter gesetzt werden.

$$arg\ min_\beta ||y - X\beta||_2^2 + \gamma_1 ||\beta||_1 + \gamma_2 ||\beta||_2^2$$

Hier werden sowohl die L1- als auch die L2-Norm minimiert.

9.2.6 Overfitting

Stichprobenverzerrung resultiert in zu optimistischen Resultaten. Feature Selection und Vorhersagen dürfen nicht mit dem selben Datenset gemacht werden. Man benutzt also einen Trainings-Datensatz und einen Test-Datensatz.

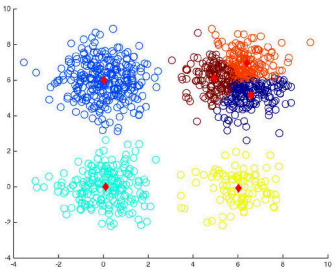
10 Clustering

Beim Clustering versucht man Gruppen von Datenpunkten zu bilden, welche einander ähnlich sind.

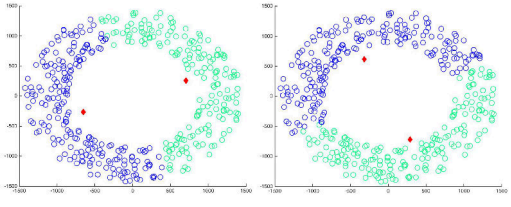
10.1 Clustering-k means, Centroid based

Jeder Cluster wird von einem Vektor repräsentiert, welcher kein existierender Datenpunkt sein muss. Man ordnet sie wie Folgt:

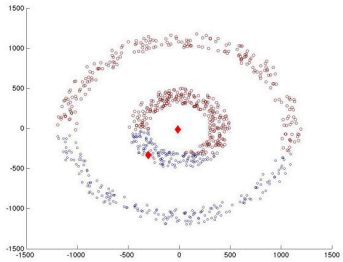
1) wähle k zufällige Cluster-Mittelpunkte
2) ordne jeden Punkt dem nächsten Mittelpunkt zu
3) berechne den neue Mittelpunkt des Clusters
4) wiederhole **2)** und **3)** bis sich nichts mehr ändert



k muss vom Benutzer gefunden werden



je nach Startpunkt können Ergebnisse unterschiedlich sein



nicht-sphärische Cluster bereiten Probleme

10.2 Graph-based Clustering

Annahmen

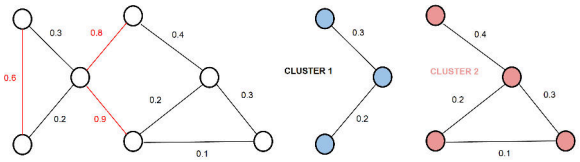
- Daten sind in Form eines Netzwerks / Graphen
- jeder Punkt ist ein Objekt
- Kanten verbinden verwandte Objekte
- Gewichtung von Kanten repräsentiert die Entfernung zwischen Objekten

Graphen stammen aus

- Literatur über biologische Netzwerke
- *treshold distance matrix*

Vorgehen

1) entferne alle Kanten, die schwerer gewichtet sind als eine bestimmte Vorgabe
2) finde alle verbliebenen Punkte → Cluster



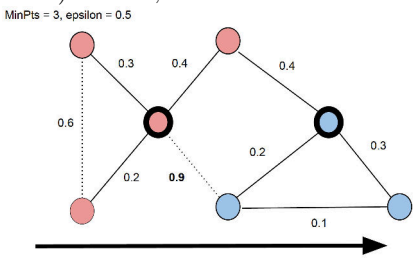
10.3 DBSCAN

Rausch-resistente Variante von Graphen-basiertem Clustering: **Density Based Spatial Clustering of Applications with Noise**. Dabei gibt es drei Arten von Punkten:

- *core object*: Punkt mit Mindestanzahl (MinPt) von Punkten in Entfernung Epsilon
- *border point*: innerhalb Epsilon eines *core objects*
- *noise*: restliche Punkte

Algorithmus

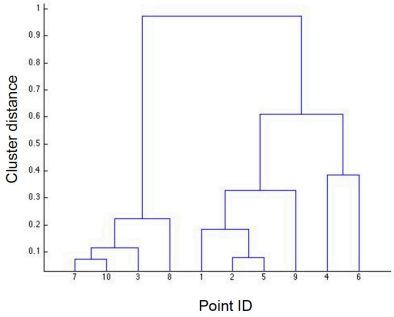
1) wähle noch nicht zugewiesenen Punkt
2) ist er ein *core object*, erstelle ein neues Cluster, sonst markiere ihn als *noise*
3) füge seine Nachbarn zum selben Cluster hinzu
4) kontrolliere jeden Nachbarn, ob er ein *core object* ist
  1) wenn ja, füge seine Nachbarn zum selben Cluster hinzu und wiederhole **4**) für diese
  2) falls nein, Cluster nicht weiter wachsen lassen
5) kehre zu **1**) zurück, bis alle Punkte erfasst wurden



Vorteile	Nachteile
kein k benötigt	zwei Parameter müssen gewählt werden
Cluster, die k-means nicht findet	Initialisierungs-abhängige Resultate
robuster gegenüber Rauschen	bei variierender Dichte (häufig bei hochdimensionalen Datensätzen) werden viele Cluster nicht gefunden

10.4 Hierarchical Clustering

In den anderen Methoden sind Cluster einander gleichgestellt. Mit dieser Methode clustert man kleinere Cluster immer wieder zu grösseren Clustern. Man Beginnt mit jedem Datenpunkt als Cluster und beendet den Prozess, wenn nur noch ein grosses Cluster übrig ist.



Vorteile	mehr Einsicht in die Struktur der Daten
Nachteil	für weitere Verwendung braucht man meist gleichgestellte Cluster

11 Clustering Applications

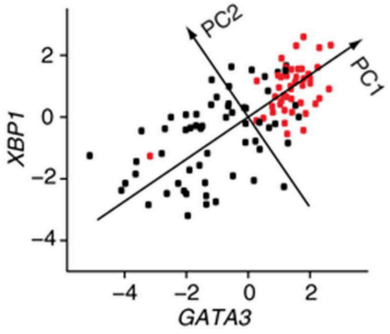
Clustering von *samples* (verschiedene Mutanten, Bedingungen, Behandlungen) über *features* hinweg erlaubt es, Zelltypen oder Ähnlichkeiten in (intrazellulären) dynamischen Antworten zu identifizieren. Z.B. Timing von Ereignissen, konservierte Antworten auf Medikamente, Proteinkomplexe aus phänotypischen Antworten in *knock-outs*.
Clustering von *features* (Gene, Metaboliten, Proteine) über *samples* hinweg erlaubt es, korrelierte *features* zu finden und ist oft mit *enrichment analysis* verbunden.
Co-clustering wird zur Rauschreduktion verwendet.

11.1 Dimensionsreduktion

In grossen Datensätzen (viele Features = viele Dimensionen) will man die Anzahl Dimensionen reduzieren, ohne dabei Informationen über die Ähnlichkeit von Samples zu verlieren. Dabei versucht man z.B. Redundancen und Rauschen zu entfernen. Weniger Dimensionen sind nützlich zur Visualisation, Qualitätskontrolle und Verarbeitung von Daten bevor andere Methoden benutzt werden.

11.2 Principal Component Analysis (PCA)

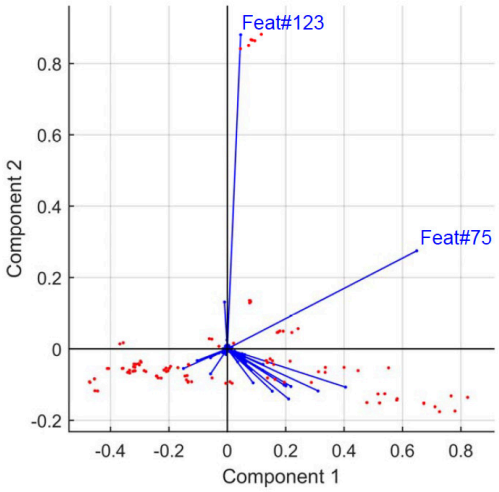
Die PCA versucht PCs in multidimensionalen Daten zu finden. PCs sind lineare, orthonormale Kombinationen der ursprünglichen Dimensionen. Sie werden dabei nacheinander so bestimmt, dass die Varianz entlang der Projektion jeweils maximal ist. Eigenwerte sind die Varianz der Punkte projeziert auf jede einzelne Komponente.



Typische Kriterien um die Anzahl der zu behaltenden Komponenten zu bestimmen sind

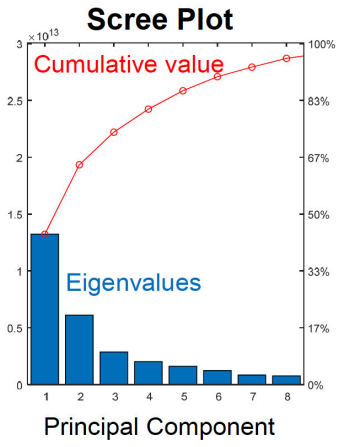
- k Komponenten erklären einen bestimmten prozentualen Anteil der Varianz (z.B. 75%)
- k Eigenwerte sind über dem durchschnittlichen Eigenwert
- Wo ist der letzte grosse Sprung im Scree-Plot (siehe 11.2.2)?

11.2.1 Bi Plot



- nahegelegene Samples haben ähnliche Profile
- nahegelegene Features korrelieren
- Samples nahe an einem Feature haben alle hohe Levels des selben Features
- Samples gegenüber voneinander haben geringe Levels des selben Features

11.2.2 Scree Plot



11.3 Recap

- Gruppen unbekannt
  - Clustering
  - Dimensionsreduktion → visuelle Inspektion
- Gruppen bekannt
  - Feature selection
    - univariate
    - multivariate

12 Link Prediction

Was ist link prediction (*collaborative filtering*)? Bei einem gegebenen Netzwerk will man fehlenden Kanten hervorsagen, z.B. um herauszufinden, welcher Transkriptionsfaktor welches Gen beeinflusst. Man brauch dazu Gen-Expressions-Daten, eine bekannte Netzwerkstruktur und eine Gen- / Protein-Sequenz.

Unsupervised link prediction: Nicht auf das Lernen aus Beispielen ausgelegt, sondern auf vordefinierten Regeln.
Supervised link prediction: Man basiert das Wissen auf bekannten Beispielen (Komponenten die interagieren oder nicht interagieren).

12.1 Similarity-Based Approach

Setze Kante zwischen Genen a und b, wenn ihre Ähnlichkeit $s(a, b)$ über einem bestimmten Wert θ liegt.

Vorteile	Nachteile
einfach umzusetzen	man muss θ setzen
skaliert zu grossen Netzwerken	Ähnlichkeit muss nicht = Interaktion sein

Typische *similarity measures* in *link prediction* sind:

- Pearsons correlation coefficient
- Mutual Information
- String kernels that count common subsequences in two protein sequences (k-mers)
- Number of shared neighbors

12.2 Cluster-Based Learning

Wenn a und b im selben Cluster liegen, sagt man eine Kante voraus.

Vorteile	Nachteile
allgemeiner als *similarity based pairs*	wie realistisch hängt von Güte der Cluster ab

12.3 Latent Group Models

Das Ziel ist es Interaktionen zwischen Genen herzuleiten. Als erstes kommt die Trainingsphase:

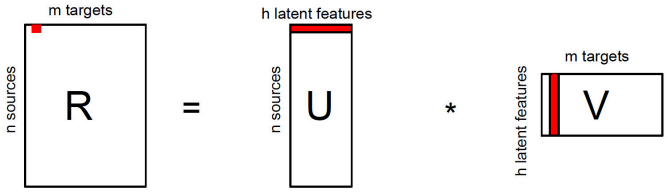
1) wähle eine Grupps S von Genen
2) cluster die Gene von S in k verschiedene Gruppen aufgrund der Expressionsprofile
3) Für jedes Paar Cluster i und j bestimme die empirische Interaktionswahrscheinlichkeit p_{ij}

Nun will man eine Voraussage für zwei Gene machen:

5) gegeben ist ein Paar Gene a und b
6) weise a und b den ähnlichsten Cluster C_a und C_b zu
7) finde die Interaktionswarscheinlichkeit $p_{a,b}$ von der Trainingsphase.

Hängt von einem Set von latenten oder versteckten Eigenschaften ab, während *cluster-based link prediction* nur von einer latenten Variable abhängig ist.

12.4 Matrix Factorization

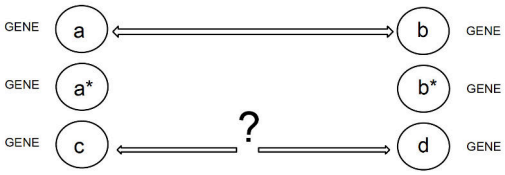


$$R_{a,b} = \sum_{i=1}^{h} U_{a,i} V_{i,b}$$

Da nicht alle Interaktionen in R bekannt sind, zerlegt man die Matrix um die restichen Werte vorauszusagen.

12.5 Kernel approaches: Similarity-based classification

Für manche Genpaare ist bekannt, dass sie interagieren und für andere das sie nicht interagieren. Wenn man nun ein Genpaar hat, über das nichts bekannt ist, versucht man sie mit anderen zu vergleichen und aufgrund der Ähnlichkeit vorherzusagen, wie sie interagieren. Kernel ist dabei eine Ähnichkeitsfunktion.



Tensor pairwise kernel bestimmt die Ähnlichkeit der beiden Knotenpaare bezüglich ihrer Kanten in beide Richtungen.

$$\begin{aligned} k_{tensor}((a,b),(c,d)) =& k_{nodes}(a,c)k_{nodes}(a,c) \\ &+ k_{nodes}(a,d)k_{nodes}(b,c) \end{aligned}$$

Metric learning pairwise kernel: $\varphi(g)$ ist ein Vektor, der die Eigenschaften vom Gen/Protein g beschreibt.

$$k_{ml}((a,b),(c,d)) = \left[(\varphi(a) - \varphi(b))^T(\varphi(c) - \varphi(d))\right]$$

Das Paar (a,b) ist ähnlich zu (c,d), wenn $a-b$ ähnlich zu $c-d$ oder $d-c$ ist.

Negativ Kontrollen gut zu wählen ist schwierig. Die häufigste Strategie ist es Proteine von anderen Zellkompartimenten auszuwählen. Oft ist das aber eine zu grosse Vereinfachung. Eine andere Strategie ist es, die positiven und negativen Kontrollen gleich zu gewichten, was jedoch zu einem zu pessimistischen Resultat führt.

12.6 Regression-based

Sage den Zustand eines Gens voraus, wenn alle anderen Genen gegeben sind. Jedes Gen ist durch einen Vektor repräsentiert (mit Expressionswerten). Man macht dann eine Regression. Meistens wird LASSO für die Regression benutzt, was die Annahme unterstützt, dads ein Gen jeweils nur mit einem kleinen Set von anderen Genen interagiert.

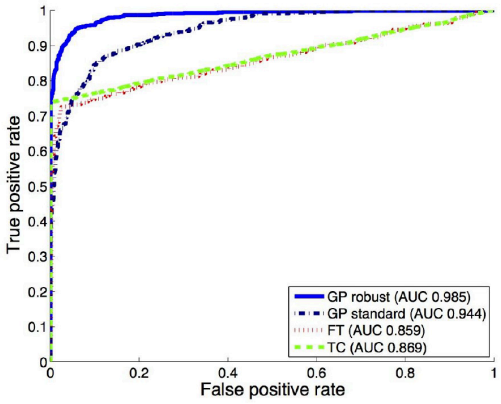
true (R) vs. predicted label (R^*)	$R=1$	$R=-1$
$R^*=1$	TP	FP
$R^*=-1$	FN	TN
	P	N

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

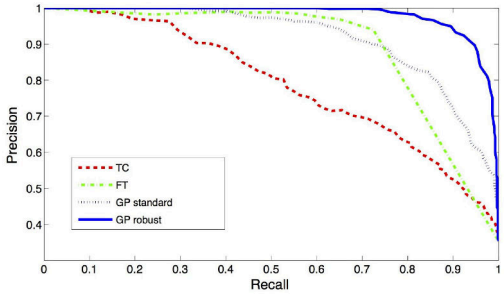
Welcher Anteil an Vorhersagen ist korrekt?

- **sensitivity** (recall / true positive rate) $= \frac{TP}{P}$. Wie hoch ist die Rate an positiven Beispielen, die der classifier fand?
- **specificity** $= \frac{TN}{N}$
- **false positive rate** $= \frac{FP}{N} = 1 - \text{specificity}$
- **precision** $= \frac{TP}{TP+FP}$. Welcher Anteil an positiven Vorhersagen ist korrekt?

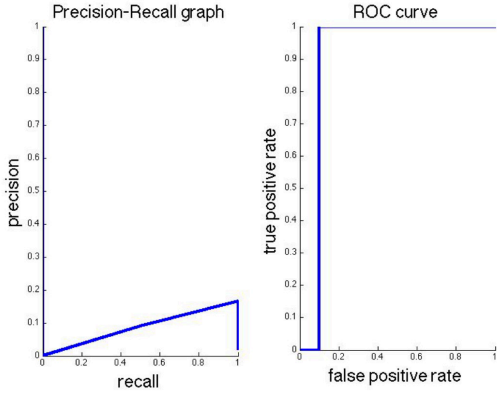
Receiver operating characteristic (ROC)



AUC = Area under the ROC curve → Ist die Wahrscheinlichkeit das positive Sample zu finden, wenn es mit einem positiven und einem negativen Beispiel konfrontiert ist. Pitfall: Auch mit einem sehr hohen AUC ist der Classifier ziemlich unnötig, wenn eine Klasse viel grösser ist als die andere. Dann muss man sich die Presicion/Recall Kurve anschauen.



Beispiel: We perform link prediciton for 102 pairs of nodes. For 2 pairs, a link actually exists. They are ranked in the 11th and 12th position of the solution.



AUC/ROC is misleading here and looks much better than the precision-recall curve.

13 Biological Networks

- **Gene regulatory networks**
  Transkriptionsfaktor-Gen-Interaktionen
- **Protein-protein interaction networks**
  Protein-Protein-Interaktionen
- **Phosphorylation networks**
  Kinase/Phosphatase-Ziel-Interaktionen
- **Metabolic networks**
  Enzym-Metabolit-Interaktionen

13.1 Rekonstruktion von biologischen Netzwerken

Experimentell kann nur ein kleiner Teil alles möglichen Interaktionen untersucht werden, da Ressourcen limitiert und Interaktionen dynamisch sind als auch Methoden sich überschneidende Resultate liefern. Trotzdem sind experimentelle Methoden weiterhin wichtig um die Basis zu legen. Rechengestüzte Methoden werden zur Hilfe genommen um das Netzwerk mit *link prediction* zu erweitern und überprüfen als auch um aus bekannten Interaktionen diejenigen zu finden, welche aktiv und wichtig für den Phänotyp sind.

13.1.1 Probleme von Vorhersagemethoden

- Schlecht darin, kurzanhaltende Interaktionen und Interaktionen mit schlecht untersuchten Interaktionspartnern vorherzusagen.
- Wenn eine Methode Training benötigt, kann sie ein *bias* für den Trainingsdatensatz erhalten.

13.1.2 Mehrere Methoden

Die Analyse von verfügbaren Methoden zeigt, dass jede Familie von Methoden bestimmte Zusammenhänge besser vorhersagt als andere. Es ist also stets besser, Methoden aus verschiedenen Familien zu kombinieren.

13.2 Nutzen von Netzwerken

- entdecken von Funktion und Organisation
- durch differenzielle Analyse dynamische oder konservierte Module finden
- Netzwerke unterstützen Dateninterpretation

13.2.1 Effekte von Medikamenten

Ein grosses Problem in der Suche nach neuen Medikamenten ist, dass bei der Zugabe eines Stoffes zu einer Zelle Hunderte bis Tausende Genprodukte direkt oder indirekt auf die Veränderungen reagieren. Eine mögliche Herangehensweise um die direkten Effekte rauszufiltern sieht wie folgt aus:

1) Zelle auf verschiedene Arten behandeln
2) RNA-Expression für jede Behandlung messen
3) Netzwerkmodell erstellen
4) Zelle mit Medikament behandeln
5) RNA-Expression messen
6) Daten mithilfe des Modells filtern → nur direkte Ziele des Medikaments bleiben übrig.

13.2.2 Krebs-Kategorisierung

Viele Krebsarten haben mehrere Unterarten mit verschiedenen Ursachen und Folgen. Tumor-Genome sind wichtige Informationsquelle, sind aber schwierig zu vergleichen, da zwei Tumore selten die selben Mutationen haben. Mit *network-based stratification* (NBS) können Tumor-Genome mit Gennetzwerken integriert werden. Dies erlaubt die identifizierung von Unterarten durch *clustering* von Mutationen in ähnlichen Netzwerkregionen. Mithilfe des Modells können für jede Unterart mRNA-Expressionssignaturen erstellt werden.