

SYSTEMBIOLOGIE

WOCHE 1

SYSTEMBIOLOGIE

In der Systembiologie versucht man mithilfe computergestützter mathematischer Modelle zu verstehen, wie das **Zusammenwirken molekularer Komponente** zum **Verhalten und Eigenschaften** biologischer Systeme führt

→ Lehre vom **System (Netzwerk-Biologie, Syntax)**

→ beschreibt das **Zusammenspiel mehrerer Komponenten**

→ versucht **komplexe und dynamische Interaktionen der Komponenten** eines biologischen Netzwerkes zu verstehen

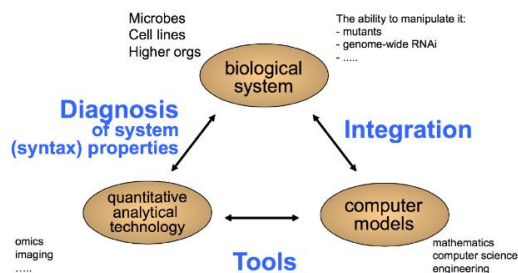


Fig. 1: Systembiologie arbeitet mit einer Kombination aus „large-scale“ Omics-Experimenten, Manipulierung des Systems und rechengestützten Analysen/Modellen.

- SB wird durch die Kombination von gross angelegten experimentellen Omics-Analysen, Manipulation biologischer Systeme und Computational Analysis/Modeling angetrieben

Schwierigkeit:

- unendliche Menge an Daten und Verknüpfungen
- Ein biologisches System ist mehr als nur die Summe seiner Teile → **emergent properties**: Netzwerk besitzt Eigenschaften, die in den einzelnen Komponenten nicht vorhanden sind (Bsp. 2 Messer + Schraube = Schere)
- Netzwerke haben keine linearen Kausalitäten

Biologische Systeme sind komplex

- grosse Anzahl Komponenten
- unterschiedliche Formen von Interaktion
- selbst-modifizierende Systeme (Störungen umgehen)
- räumliche und zeitliche Organisation (Trennung von Teilprozessen)
- häufig noch unbekannte Komponenten und Interaktionen (Umwelt als Interaktionskomponente)
- Es gibt keine Korrelation zwischen Komplexität eines Organismus und dessen Anzahl Gene (Ein Gen codiert nicht für ein Protein mit einer einzigen Wirkung!)

EINSCHRÄNKUNG GENETISCHER ANSÄTZE

- Komplexität korreliert nicht direkt mit Anzahl Genen
- Medikamente wirken auf ein Teil eines Netzwerkes → Mehrere Effekte und Nebeneffekte
- Zellantwort hängt von vielen Faktoren ab → nicht nur genetisch, sondern auch Umwelt bestimmt

DIFFERENTIALGLEICHUNGEN/ ODEs

Dynamisches Modell (ODE = ordinary differential equations)

- Für **dynamische Prozesse** verwendet man immer ODEs!
- Gleichungen, wo man die Ableitung einer Grösse durch eine Funktion der Grösse selbst ausdrückt
- Differentialgleichungen verwendet man immer dann, wenn der aktuelle Zustand eines Systems seine weitere Entwicklung beeinflusst

ANNAHMEN/BEDINGUNGEN:

- Grosse Anzahl an Molekülen
- Gut durchmisches System mit sich frei bewegenden Metaboliten (keine Heterogenität)

BENÖTIGTE INFORMATIONEN:

- Komponente und Konzentrationen
- Kinetische Parameter
- Massen- und Energiebalancen
- Thermodynamik

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{p}, \mathbf{u}(t)) \quad , \quad \mathbf{x}(t_0) = \mathbf{x}_0$$

- System of ordinary, first-order, linear or nonlinear differential equations (ODEs) characterized by:

- Right hand sides $\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p})$ = **reaction rates**.
- System states $\mathbf{x}(t)$ = **dynamic concentrations**.
- Parameters $\mathbf{p}$ = **kinetic (rate) constants**.
- Inputs $\mathbf{u}(t)$ = **external manipulations**.
- Initial conditions $\mathbf{x}(t_0)$ = **initial concentrations**.

Beispiel: Das Wachstum einer Zelle hängt oft davon ab, wie gross die Zelle zurzeit ist. Mathematisch bedeutet das, dass die **zeitliche Ableitung (=Veränderung) des Zellvolumens** $\frac{dV}{dt}$ eine Funktion des momentanen Volumens V ist: $\frac{dV}{dt} = f(V)$

MASSENWIRKUNGSGESETZ

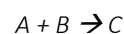
= *mass action law*

Geschwindigkeit einer chemischen Reaktion ist **proportional zum Produkt der Konzentrationen der Reaktanden**

- zeitliche Konzentrationsveränderung der Stoffe in Abhängigkeit ihrer aktuellen Konzentration
- biochemische Kinetik basiert auf dem Massenwirkungsgesetz
- Die **Michaelis-Menten-Kinetik** ist ein Spezialfall der Massenwirkung für enzymkatalysierte Reaktionen mit Ratenkonstanten für jede Teilreaktion → zusätzliche Annahmen

BEDINGUNGEN:

- bei konstanter Temperatur
- ohne Katalysator



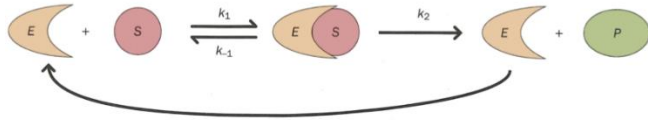
$$\frac{d[A]}{dt} = -k[A][B], \quad \frac{d[B]}{dt} = -k[A][B] \quad \text{und} \quad \frac{d[C]}{dt} = +k[A][B].$$

Die Konzentration von A verändert sich mit einer Rate, die proportional zur Konzentration von A und B ist, da diese beiden Reaktanden zusammenkommen müssen, um C zu bilden

MICHAELIS-MENTEN-KINETIK

→ Chem. Reaktionen, die durch Enzyme katalysiert werden

Michaelis-Menten-Gl. beschreibt die **Geschwindigkeit** solch einer **enzymatisch katalysierten Reaktion**



$$v = \frac{v_{\max} \cdot [S]_0}{[S]_0 + K_M}$$

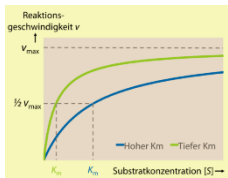
v = Reaktionsgeschwindigkeit
 $v_{\max}$ = maximale Geschwindigkeit, die die Reaktion erreichen kann

$$v_{\max} = k_2 \cdot [E]_{\text{tot}}$$

$k_2 = k_{\text{cat}}$ = Anzahl Substratmoleküle, die ein Enzym pro Sek. zu Produkt umsetzen kann

$$K_M = \frac{k_{-1} + k_2}{k_1}$$

K_M = Michaelis-Menten-Konstante: Substratkonzentration, bei der die **halbe Maximalgeschwindigkeit** erreicht ist → Mass für die **Affinität** von Substrat zu Enzym → Je kleiner K_M , desto höher die Affinität und desto weniger Substrat wird benötigt, um sich $v_{\max}$ anzunähern



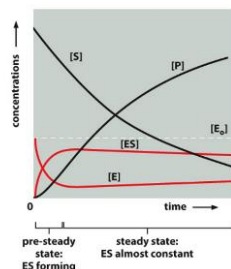
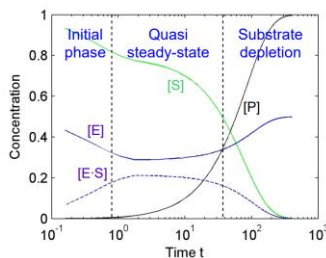
STEADY-STATE-BEDINGUNG

Steady-state = Gleichgewichtszustand, in dem alle Variablen, die das System beschreiben, gleichbleiben (konstant)

- Bedeutet nicht, dass alles stillsteht, sondern dass sich alle Ein- und Ausströme zu einer Variable ausgleichen
- **Fliessgleichgewicht**

Quasie-steady state=

- Annäherung an einen Endzustand
- Phase, in der die ES-Konzentration näherungsweise konstant bleibt ($[ES] = \text{konstant}$)
- Fliessgleichgewicht
- Kann aber trotzdem noch oszillieren oder sogar chaotisch verbleiben



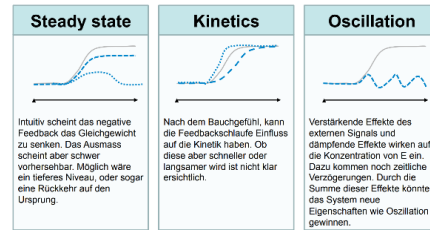
Die steady-state-Annahme (Pseudo-steady-state) vereinfacht die mathematische Modellierung deutlich, weil durch die zeitliche Konstanz aus Differentialgleichungen algebraische Gleichungssysteme werden

Zelle stellt Fliessgleichgewichte her durch z.B.

- Proteinabbau
- ständige Teilung → Verdünnung durch das Wachstum
- negatives Feedback

NEGATIVES FEEDBACK

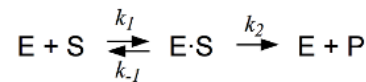
Kann **Steady State** und **Kinetik** auf nicht vorhersehbare Art beeinflussen



HERLEITUNG DER MICHAELIS MENTEN GLEICHUNG

1. ODES (DYNAMISCH) DER **MASS-ACTION KINETIK** ERSTELLEN

(für die Herleitung der Michealis-Menten Kinetik)



$$\frac{d[S]}{dt} = -k_1[E][S] + k_{-1}[E \cdot S]$$

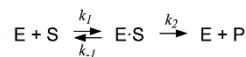
$$\frac{d[E]}{dt} = -k_1[E][S] + (k_{-1} + k_2)[E \cdot S]$$

$$\frac{d[P]}{dt} = k_2[E \cdot S] = v$$

$$\frac{d[E \cdot S]}{dt} = +k_1[E][S] - (k_{-1} + k_2)[E \cdot S]$$

Wenn die Reaktion etwas bildet, nehme man die **Rate (+) mal** die **Konzentrationen** jener, welche dieses bilden

2. VEREINFACHUNGEN



1. **Keine Rückreaktion** vom Produkt und **kein Feedback** des Produkts (irreversible Reaktion)
2. Totale **Enzymkonzentration bleibt konstant**
 → $[E]_{\text{tot}} = [E] + [E \cdot S]$ → Einsetzen in $\frac{d[S]}{dt}$ und $\frac{d[ES]}{dt}$

3. ANNAHMEN

3. Quasi **steady-state**: (Bildung und Dissoziation viel schneller als Produktbildung & **konstante Konzentration des Enzym-Substrat-Komplexes [ES]**)
 → Summe der Einströme (/Umwandlungen) zu ES = Summe der Ausströme von ES
 → $k_1, k_{-1} \gg k_2$: $\frac{d[ES]}{dt} = 0$
4. **Substratüberschuss** $[S] = [S]_0$

4. ALLES IN D[P] EINSETZEN:

$$[E \cdot S] = \frac{[S]_0 [E]^T}{[S]_0 + \frac{k_2 + k_{-1}}{k_1}}$$

$$\frac{d[P]}{dt} = +k_2 [E \cdot S] = \frac{k_2 [S]_0 [E]^T}{[S]_0 + \frac{k_2 + k_{-1}}{k_1}} \equiv v$$

Die Annahmen bei der Herleitung der Michaelis-Menten-Gleichung sind auf andere Geschwindigkeitsgesetze von biologischen Phänomenen übertragbar

Transkriptionsfaktor T bindet an eine DNA Sequenz G :
 Michaelis-Menten-Konstante K gibt die Affinität des Transkriptionsfaktors für die DNA Sequenz G mit der totalen Konzentration $[G]^T$ an

- Without repression: $[G \cdot T] = \frac{[G]^T [T]}{[T] + K}$
- Competitive repressor R : $[G \cdot T] = \frac{[G]^T [T]}{[T] + K(1 + [R]/K_I)}$
- Cooperative binding: $[G \cdot T] = \frac{[G]^T [T]^n}{[T]^n + K^n}$

MODELLE

Zum Generieren einer begrifflichen, physischen oder mathematischen Darstellung eines realen Phänomens, das schwierig zu beobachten ist

Topologische Modelle: Abstraktion der Realität, Aspekte der Wirklichkeit werden **vereinfacht**

- "All models are wrong, but some models are useful."

Zur Interpretation von Daten und Hypothesen benötigt man häufig formale Modelle, da sogar „einfache“ biologische Systeme ein unintuitives Verhalten zeigen können.

- (Wachstum hat Einfluss auf die Genexpression, da das Zellvolumen und somit die Konzentration verändert wird)
- Man kann mathematische Modelle nutzen, um biologische Systeme zu beschreiben (Werkzeugkasten)
- Meistens gibt es mehrere Erklärungen für ein Verhalten
→ man muss die verschiedenen Möglichkeiten kritisch hinterfragen

REDUKTIONEN DURCH MODELLE

Modelle als **Vereinfachung**

→ Ausgrenzung eines Teils der Realität!

Entscheiden, wo man die Grenzen zieht → **Kompromiss!**

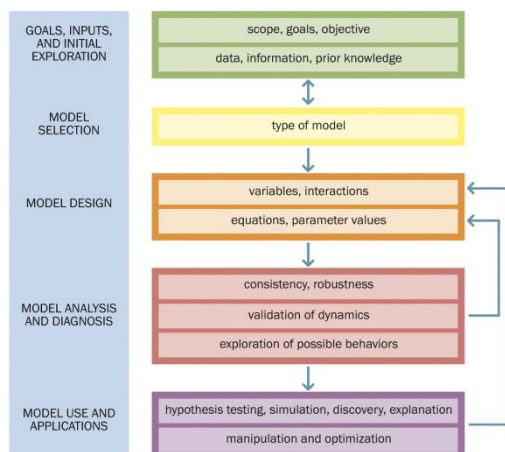
- Zeit, Dynamik, Verzögerung
- Interaktionen, Crosstalk → Netzwerk reduzieren
- Interaktion mit der Umwelt, Signal
- Output oft nicht genau spezifiziert

Problem bei der Modellierung: Konzeptionelle Vereinfachung, Reduktion, pragmatischen Umgang

Ist diese Vereinfachung noch hinreichend, um das System zu verstehen? "Everything should be made as simple as possible, but not simpler." (Albert Einstein)

MODELLDESIGN

Ein gutes Modell kann nicht nur eine gemachte **Beobachtung erklären**, sondern erlaubt auch die **Formulierung neuer Hypothesen** und kann **Vorhersagen** über ungetestete Situationen machen



Wissenschaftliches Modell erstellen:

1. Fragestellung und Detailebene wählen
2. Technologie: Experiment designen, Geräte entwickeln
3. Genomics: Durchführen des „wet“-Experiments, Daten Sammeln und mit biologischen Erkenntnissen zusammenführen
4. Computation: Modellieren der Daten, Stimulieren der „dry“-Experimente
5. Systemanalyse, Theoriebildung, Voraussagen

1. FRAGESTELLUNG

Nicht der modellierte Prozess selbst, sondern die **Fragestellung** über diesen Prozess bestimmt den Modellierungsvorgang

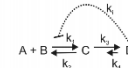
- Formulierung des Ziels
- Auswahl der richtigen Skalen für das Modell (räumlich und zeitlich)
- welche Informationen benötigt man, um die untersuchte Fragestellung zu beantworten?
- erste Erkundung der verfügbaren Informationen

2. MODELLWAHL

- Welche Modelle gibt es? (siehe 'Typen von Modellen') und welches wählt man?

Hängt das modellierte Phänomen von der Zeit ab?

- **steady-state?**
- **dynamisches Modell** falls die verwendeten Daten Veränderungen mit der Zeit zeigen → zeitliche Veränderungen spielen eine Rolle
- **statische Modelle** falls zeitliche Veränderungen keine wichtige Rolle spielen → zeitunabhängig
- **korrelative Modell**: beschreibt die einfache Beziehung zwischen zwei Merkmalen
- **erklärendes Modell**: bringt eine Beobachtung mit dem biologischen Prozess oder Mechanismus, welcher dem Phänomen zugrunde liegt, in Verbindung

Model class	Level of abstraction	Required information	Example applications
Topological (steady state)	Interaction 	Components and unspecified connections must be known	- Genetic networks - Protein-protein interaction - Metabolite-protein interaction
Stoichiometric (steady state)	Reaction stoichiometry $A + B \rightleftharpoons C \rightarrow D$	Mass and energy balances thermodynamics (directionality)	Metabolic networks - flux balance analysis - elementary flux modes -
Mechanistic (dynamic)	Enzyme mechanism and regulation 	Kinetic parameters	Kinetic models (including regulation)
Steady state			
Dynamic			

3. MODELLDESIGN

- Festlegen einzelner Komponente des Modells und deren Interaktionen
- **Modelstruktur/ Topologie**: Wie interagieren diese Komponenten miteinander?
- Zusammen mit dem Input bestimmen die Modellkomponenten und deren Interaktionen den Output des Modells

System state: Momentaufnahme, in welchem Zustand sich das System befindet → Werte der Variablen zu best. Zeitpunkt

Model Behavior: Veränderung des Systems über die Zeit Abhängig von:

- Umwelt (was geht hinein und hinaus, Input)
- interne Prozesse (Reaktionen, Interaktionen)

Verschiedene Modelle führen manchmal trotzdem zu derselben Lösung, warum? → Daten sind nicht hinreichend, um zwischen den Hypothesen/Alternativen zu unterscheiden/ zu differenzieren!

KLASSEN VON MODELLKOMPONENTEN:

1. **Zustandsgrößen (state variables):** repräsentieren den Zustand biologischer Objekte wie Genen, Zellen oder Molekülen und beschreiben das Verhalten des Modells
2. **abhängige Variablen:** ihr Wert ist veränderlich und sie sind abhängig von anderen Komponenten des Modells (diese Variablen kann man oft nicht direkt beeinflussen)
3. **Parameter:** sind festgelegte Werte, die man entweder experimentell ermitteln oder abschätzen muss
4. **Konstanten**= Parameter mit festgelegten Zahlen

- $V_{\max}$, K_m = Parameter
- S = Variable
- V = Output

$$V = V_{\max} \cdot S / (K_m + S)$$

ARTEN VON KOMPONENTEN:

- Proteine
- Gene
- RNAs
- Metabolite (Lipide etc.)

„**Omics**“= alle Komponenten einer Spezies in einem Level

Bsp.: Proteom = gesamte Anzahl an Proteinen

ARTEN VON INTERAKTIONEN:

- Sehr vielfältig!

Beispiele: Genexpression, Enzymreaktionen, Modifikationen, Aktivierungen, Phosphorylierungen etc.

Eine Komponente kann auch eine Interaktion sein!

FALLS EIN MODELL DIE GEMESSENEN DATEN NICHT BESCHREIBEN KANN:

- Anzahl der Parameter reduzieren
- Mehr Annahmen, Einschränkungen einführen

REAKTIONSNETZWERKE – DYNAMISCHE MODELLE

Biochemische Netzwerke sind die Summe vieler Elementarreaktionen. Die Konzepte der Modellierung der Elementarreaktionen sind deshalb direkt übertragbar auf die Modellierung ganzer Netzwerke

- Biochemische Reaktionsnetzwerke sind ein Vermittler bei der Umsetzung von pathway maps in dynamische Modelle
- Man kann dynamische Modelle entwickeln, indem man (elementare oder enzymatische) Geschwindigkeitsgesetze für alle Reaktionen zusammenstellt

REAKTIONSORDNUNGEN

Reaktionsordnung: Anzahl Teilchen, die bei einer Elementarreaktion zusammenkommen müssen, damit die Reaktion abläuft → abhängig von der Reaktionsordnung nimmt die Reaktionsgeschwindigkeit einen unterschiedlichen Verlauf:

0. Ordnung	unabhängig von der Edukt-Konzentration und konstante Geschwindigkeit
1. Ordnung	Geschwindigkeit ist von der Konzentration eines Edukts abhängig
2. Ordnung	Geschwindigkeit ist von zwei Eduktkonzentrationen abhängig

ELEMENTAR REAKTIONEN & DEREN ODE MODELS

Ein Netzwerk ist im Wesentlichen die Summe von vielen elementaren Reaktionen → Man kann eine Gesamtreaktion (auch Reaktionsnetzwerk) in seine Elementarreaktionen zerlegen

Konstante Produktion z.B.: Alkohol Dehydrogenase	$\xrightarrow{k_1} A$ $\frac{d[A]}{dt} = k_1$ $[A] = [A]_0 + k_1 \cdot t$	Same parameter, different initial concentrations
Linearer Abbau z.B. Alkoholabbau	$A \xrightarrow{k_1}$ $\frac{d[A]}{dt} = -k_1[A]$ $[A] = [A]_0 \cdot e^{-k_1 \cdot t}$ $t_{1/2} = \frac{\ln(2)}{k_1}$	
Autokatalyse	$\xrightarrow{k_1 \cdot [A]} A$ $[A] = [A]_0 \cdot e^{+k_1 \cdot t}$ $t_{1/2} = \frac{\ln(2)}{k_1}$	
Dimerisierung steady-state durch Anfangskonzentrationen ab!	$A + B \xrightarrow{k_1} A \cdot B$ $\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_1[A][B]$ $\frac{d[A \cdot B]}{dt} = +k_1[A][B]$ $\frac{dA}{dt} = -k[A][B]$	
Reversible Dimerisierung	$A + B \xrightleftharpoons[k_{-1}]{k_1} A \cdot B$ $\frac{d[A \cdot B]}{dt} = +k_1[A][B] - k_{-1}[A \cdot B]$ $\frac{d[A \cdot B]}{dt} = 0 \Rightarrow [A \cdot B]_{ss} = \frac{k_1}{k_{-1}}$ Dissociation constant $K_D = \frac{k_{-1}}{k_1}$	
Monomolekulare Konversion	$A \xrightarrow{k_1} B$	
Bimolekulare Konversion	$A + B \xrightarrow{k_1} C$	
Produktion und Degradation (k_2) Bsp.: Kontrollierte Genexpression und Abbau von mRNA	$\xrightarrow{k_1 u} A \xrightarrow{k_2}$ Produktion wird durch Faktor u reguliert Geschwindigkeitsgesetz: $\frac{d[A]}{dt} = k_1 u - k_2[A]$ Steady-state: $\frac{d[A]}{dt} = 0 \rightarrow [A]_{ss} = \frac{k_1 u}{k_2}$	Wie beeinflussen die Parameter die Dynamik? → Dynamik wird nur von der Degradation beeinflusst

MODELLIERUNG VON STOFFWECHSELWEGEN

Ein Stoffwechselweg besteht aus mehreren biochemischen Elementarreaktionen → da nun aber viele Komponenten beteiligt sind und diese netzartig miteinander verknüpft sind, zeigen sich **komplexe Verhaltensmuster**

PARAMETER

- Sind **festgelegte Werte**
- Müssen **experimentell bestimmt oder abgeschätzt** werden
- minimale Veränderungen der Parameter können zu einem völlig anderen Verhalten führen

→ Beim Modellieren ist es sehr wichtig, die Parameter richtig abzuschätzen, da das vorhergesagte Verhalten des Modells je nach Parameterwerte stark variieren kann

Parameter auswählen:

1. Wert einer Datenbank entnehmen
2. Parameter experimentell bestimmen, indem man einzelne Komponenten des Systems isoliert und die Parameter direkt misst
3. experimentelle Daten des gesamten Systems verwenden, um Parameter abzuschätzen
4. → Man muss dabei beachten, dass alle Messungen mit **Messfehler verbunden** sind

→ Ideal: Wert aus Datenbank, exakter Parameter

→ Realität: Messfehler und Unsicherheiten

→ Biologische Realität: Modelle um Parameter abzuschätzen

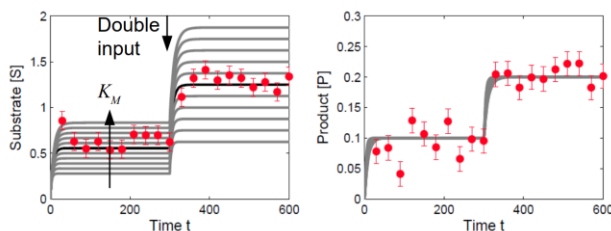
→ Schwierigkeit bei der experimentellen Bestimmung eines Parameters ist, dass man diesen oft nicht direkt messen kann, sondern anhand einer anderen Grösse ablesen muss

MESSFEHLER

- **Systematische Abweichung:** Messwerte liegen um eine gewisse **Verschiebung** neben dem wahren Wert → solche Abweichungen kann man komplett vermeiden!
- **Zufällige Messabweichungen:** unvermeidbare **Streuungen** um den wahren Wert, durch welche Messergebnisse eine bestimmte Unsicherheit haben → Die zufällige Messunsicherheit kann durch Berechnung des Durchschnitts aus einer grösseren Anzahl Messungen (grössere Stichprobe) reduziert werden

BEISPIEL KM-WERT BESTIMMUNG

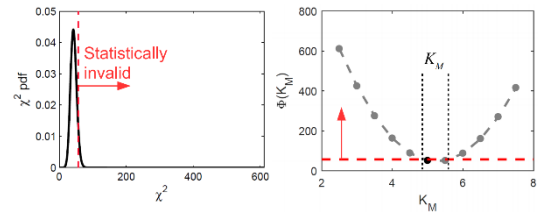
1. Modell mit unterschiedlichen Parameterwerten erstellen
2. Experimentell gesammelte Daten mit dem Modell vergleichen und jenen Parameterwert wählen, bei welchem der Unterschied am geringsten ist → gute Parameter zeigen eine kleine Distanz



[S]-Werte sind informativer für eine K_M Bestimmung als [P]

3. Unsicherheit Charakterisieren → χ^2 -test durchführen und Parameterwert nehmen, welcher im Vertrauensintervall liegt

Links: alles rechts von der roten Linie ist nicht gültig, rechts: alle Daten darüber sind nicht gültig!



PSEUDOCODES

Beispiel für einen Pseudocode:

```

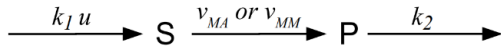
Inputs:      Dynamic pathway model
                  Experimental data
                  Possible parameter values
Output:      Objective function value for each possible
                  parameter value

For each possible parameter value
    Run dynamic simulation with current parameters
    Set objective function value = 0
    For all time points
        Compute distance between simulation and
        experimental data for time point
        Add value to objective function
    End
    Save objective function value for current parameters
End
  
```

EINFLUSS DER MODELLWAHL:

Wahl des Modells kann einen grossen Einfluss auf das vorhergesagte Verhalten des Systems haben → Welche Kinetik man annimmt, beeinflusst die Voraussage der Dynamik eines Systems (gilt bereits bei einfachen linearen Reaktionssystemen)

ENZYMKATALYSIERTE REAKTIONEN



v_{MA}: Mass-action Kinetik: Geschwindigkeit der Reaktion ist proportional zur Substratkonzentration: $v_{MA}([S]) = v_{\max}[S]$

v_{MM}: Michaelis-Menten Kinetik: $v_{MM}([S]) = \frac{v_{\max}[S]}{[S] + K_M}$

1. Substratbildung (k_1 =Produktionsrate (konst.), u =Inputfaktor)
2. Substratumsetzung → enzymatische Katalyse kann durch 2 unterschiedliche kinetische Modelle dargestellt werden: Michaelis-Menten-K oder mass-action-Kinetik
3. Produktabbau, Degradation → linearer Abbau! (Massenwirkungsgesetz)

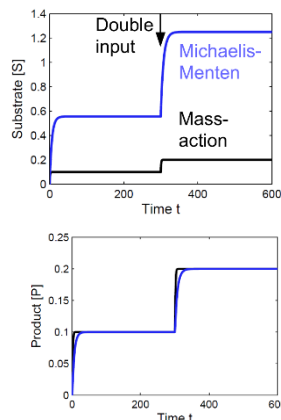
Es gilt, dass im steady state der Input in das System (Zufluss zu S) und der Output aus dem System (Abfluss von P) gleich gross und unabhängig von dem für die mittlere Reaktion gewählten Modell sind!

FRAGE 1: Hat die Wahl des Modells für die Kinetik der mittleren Reaktion einen Einfluss auf die Konzentrationen von Substrat und Produkt?

→ die Wahl des Kinetik-Modells beeinflusst stark die Vorhersage der Substratkonzentration! → Steady-state-Konzentration des Substrats ist bei MM viel höher!

→ Steady-state der Produkt-Konzentration hängt nur von k_1 , k_2 und u ab und nicht vom steady-state der Substratkonzentration oder der Wahl der Enzym Kinetik für die Substrate, warum?

→ Produktkonzentration im steady-state ist 100% Kinetik-unabhängig → Ist nur abhängig vom Influx und Efflux



Frage 2: Wie reagieren die Konzentrationen von Substrat und Produkt auf eine Veränderung (hier: Verdoppelung) des Inputs? Steigen die GGW-Konzentrationen von S und P linear mit der Reaktionsgeschwindigkeit?

→ Steady-state-Konzentration des Substrats steigt unter Annahme der MM-Kinetik **nicht linear mit der Verdoppelung des Inputs u auf das Doppelte an**, sondern um mehr als das

ERKLÄRUNG: Michaelis-Menten-Kinetik berücksichtigt die Affinität von Enzym und Substrat. Mit steigender Substratkonzentration steigt auch die Rate der Katalyse

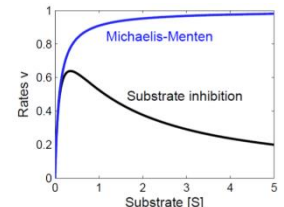
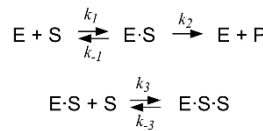
→ Im Gleichgewicht gilt: Je höher K_M -Wert des Enzyms, desto geringer die Affinität von Enzym und Substrat und desto höher die steady-state-Konzentration des Substrats.

FAZIT: Bei enzymkatalysierten Reaktionen ist die Annahme der Michaelis-Menten-Kinetik eine bessere Approximation

→ Affinität des Enzyms für das Substrat wird in Form des K_M -Werts berücksichtigt!

SUBSTRATE INHIBITION

- **NORMAL:** Geschwindigkeit einer enzymkatalysierten Reaktion **erhöht sich mit steigender Substratkonzentration** und nähert sich immer mehr $v_{\max}$ an
- **ENZYME MIT SUBSTRAT INHIBITION:** Steigende Substratkonzentration verursacht einen Rückgang der Produktbildung → Verlangsamung der Reaktion ab einem gewissen Punkt/ einer gewissen $[S]$
- Bei ca. 20% aller Enzyme



$$\text{inhibition constant } K_I = \frac{k_{-3}}{k_3}$$

Abbildung 2.6: Zwei Arten von Enzymkinetik: Michaelis-Menten-Kinetik und Substratinhibition. Bei der Substratinhibition sinkt die Reaktionsgeschwindigkeit v ab einer gewissen Substratkonzentration wieder.

$$v([S]) = \frac{v_{\max}[S]}{[S] + K_M + \frac{[S]^2}{K_I}}$$

Inhibition term has effect at high substrate concentrations

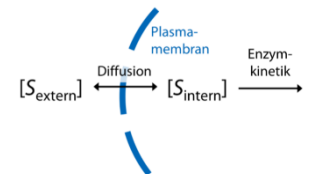
SCHALTERARTIGE KONZENTRATIONSÄNDERUNG:

Emergent behavior, hier: «switching»

Erste Reaktion: Diffusion des Substrats durch die Membran:

Geschwindigkeit ist proportional zu dem Konzentrationsgefälle auf den beiden Seiten der Membran

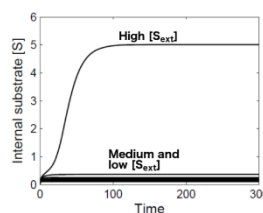
$$\rightarrow v_{\text{Transport}} = D \cdot ([S_{\text{extern}}] - [S_{\text{intern}}])$$



Zweite Reaktion: Weiterreagieren des Substrats_{intern} via enzymkatalysierter Reaktion:

$$v([S]) = \frac{v_{\max}[S]}{[S] + K_M + \frac{[S]^2}{K_I}}$$

→ Weiterreaktion durch Enzymkinetik mit Substratinhibition



BEOBACHTUNG: Man misst entweder sehr hohe oder sehr tiefe interne Substrat-konzentrationen, aber keine Werte dazwischen → Schalterartiges Verhalten eines Systems

BEGRÜNDUNG:

Im GGW gilt Einstrom=Ausstrom → Gleichgewichtskonzentration des Substrats, die sich einstellen wird, kann man am Schnittpunkt der beiden Kurven ablesen → Resultat aufgrund der **Kurvenform der Inhibitionskinetik (schwarz)**

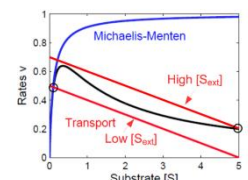


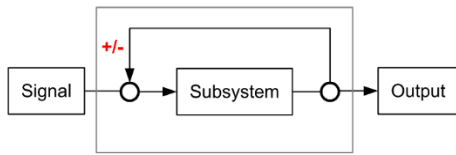
Abbildung 2.9: Grafische Begründung für das schalterartige Verhalten des Reaktionssystems aus Diffusion (rot) und Enzymkinetik bei Substratinhibition (schwarz). Für hohe externe Substratkonzentrationen (obere rote Gerade) liegt der Schnittpunkt mit der Kurve für die Enzymkinetik bei einer hohen internen Substratkonzentration. Bei mittleren oder tiefen externen Konzentrationen (untere rote Gerade) stellt sich ein Gleichgewicht bei einer sehr tiefen internen Substratkonzentration ein.

Emergentes Verhalten wie "Schalten" kann auch bei zwei gekoppelten Reaktionen auftreten

W3) REGULATIONSMECHANISMEN

RÜCKKOPPLUNGSMECHANISMEN (FEEDBACK LOOPS)

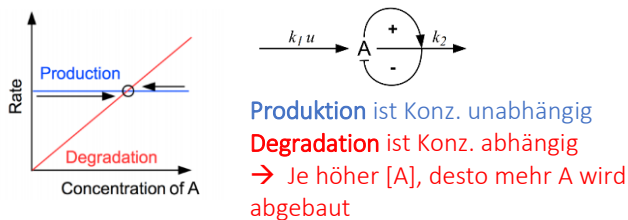
Feedbackmechanismen werden benötigt, um den unbegrenzten Anstieg eines Stoffes zu verhindern. Feedbackloops führen dazu, dass der Output eines Prozesses den Prozess selbst reguliert:



- **negativen Feedbackloops:** Produkt inhibiert seine eigene Produktion
- **positive Feedbackloops:** Output stimuliert seine eigene Produktion (Bsp. *Caspase-Apoptose Reaktion*)
- Flüsse in einem biologischen Pathway müssen **regulierbar** sein, um **ungewollte Verhaltensweisen wie overshoots zu verhindern**

NEGATIVE RÜCKKOPPLUNG

BEISPIEL:



Schnittpunkt = angestrebtes GGW = steady state

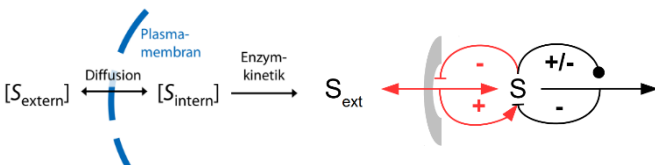
FUNKTION DES NEGATIVEN FEEDBACKSYSTEMS:

1. Einstellen eines bestimmten Gleichgewichtszustands nach einer Abweichung davon (**Homöostase**)
2. Zeit bis zur Erreichung des GGW deutlich verkürzen

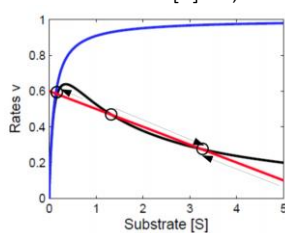
Durch negative Feedbacksysteme stellt sich schnell ein bestimmter GGW-Zustand ein (es gibt nur 1 steady state)

POSITIVES FEEDBACKSYSTEM

BEISPIEL:



1. Abhängig von der externen Substratkonzentration stellt sich „schalterartig“ ein steady state für die innere Substratkonzentration ein
2. Bei genau derselben externen Konzentration können sich **zwei verschiedene steady states einstellen** → hängt von internen [S] ab, welches GGW sich einstellt!



Bsp.: Substratinhibition

URSACHE: **Gerade**, hat bei manchen Werten für $[S_{\text{extern}}]$ mehrere Schnittpunkte mit der **Kurve für die Enzymkinetik** → Abhängig von vorheriger interner Substratkonzentration

GEDÄCHTNISFUNKTION: Abhängig von einem früheren Zustand (hier: der internen Substratkonzentration) stellt sich bei genau denselben Bedingungen ein anderer Gleichgewichtszustand ein

FUNKTION EINES POSITIVEN FEEDBACKSYSTEMS:

1. erlauben das Einstellen **mehrerer unterschiedlicher steady states** (hängt von Parameterwerten ab) → erlaubt das Switchen und Gedächtnisfunktionen
2. können kontinuierliche Signale in diskrete Zustände überführen
3. In nicht-linearen Systemen ermöglichen sie eine **Gedächtnisfunktion**
4. Möglicherweise irreversible Entscheidungen (Entwicklung) (erreicht man durch eine Verschiebung der Konzentrationen)

Durch positive Feedbacksysteme können kontinuierliche Inputsignalen in verschiedene diskrete Gleichgewichtszustände überführt werden (memory)

BIDIREKTIONALE PATHWAYS

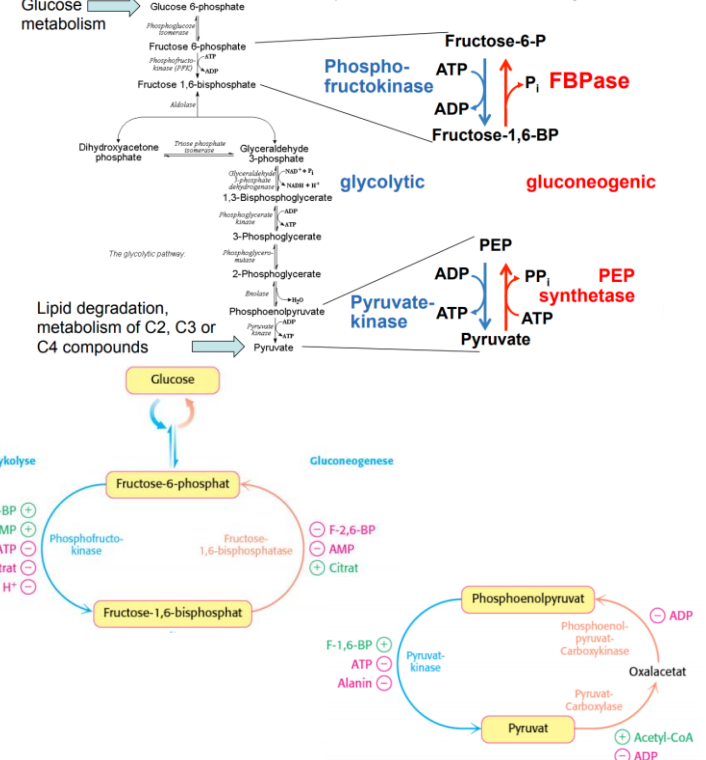
- Pathways, die in 2 Richtungen laufen müssen
- Beispiel: *Glykolyse und Gluconeogenese*

HERAUSFORDERUNGEN

1. **Schneller Wechsel der Richtung**
Mindestmenge an Enzymen und auch notwendige Cofaktoren müssen stets **vorhanden** sein, damit der schnelle Richtungswechsel möglich ist
2. **Vermeiden von futile cycles («nutzloser Zyklus»)**
Energieverschwendung dadurch, dass beide Pathways (katabol und anabol) **gleichzeitig aktiv** sind → kann durch Regulationsmechanismen vermieden werden

Regulationsmechanismen in bidirektionalen Pathways erlauben einen schnellen Richtungswechsel und verhindern schädliches Verhalten wie z.B. futile cycles

Except for 2, all reactions are freely reversible



FOLGEN EINER FEHLENDEN REGULATION

Was passiert, wenn ein Pathway nicht reguliert wird?

- Lange Zeit, bis sich ein GGW-Zustand einstellt
- Overshoot von Metabolitenkonzentration

REGULATIONS-HERAUSFORDERUNGEN

Die Erzielung angemessener dynamischer Reaktionen auf Störungen ist ein generelles Kontrollproblem für alle biologischen Systeme

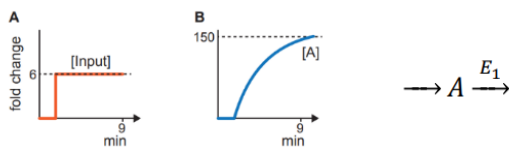
Dadurch, dass die Zelle:

- Starke Antworten verhindern möchte
- möglichst schnell wieder ins steady-state kommen will
- flexibel auf Umweltveränderungen reagieren möchte
- trotzdem Robust bleiben möchte → bei kleinen Schwankungen nicht zu schnell ändern

OVERSHOOT

→ Bei einer Erhöhung des Inputs geht die Produktkonzentration weit über den Anstieg des Inputs hinaus → stark erhöhte Metabolitenkonzentration

[A] = Produktkonzentration

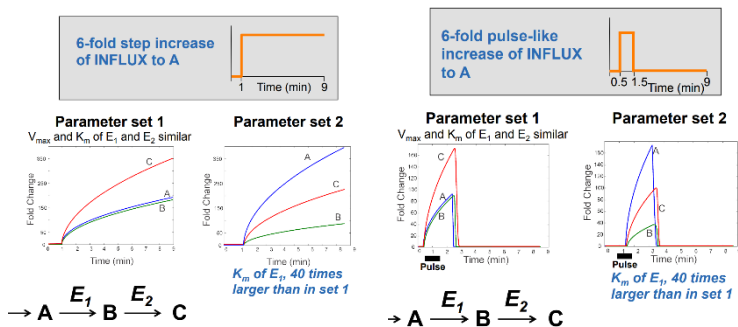


KONSEQUENZEN:

- [A] steigt unbegrenzt an
- [A] erreicht einen Wert, der unverhältnismässig grösser ist als die Steigerung des Inputs
- Es kann sich **kein neuer steady-state** einstellen
- Gravierende Folgen für eine Zelle

FOLGEN EINES OVERSHOOTS:

- Lange Zeit, bis sich ein Steady-state einstellt
- Overshoot von Metaboliten [Konzentrationen]



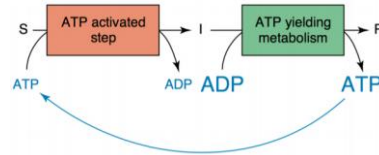
POTENTIELLE PROBLEME:

- Toxizität
- Ressourcenverschwendung
- Energieverbrauch (ATP)
- Störung anderer Reaktionen (wo Metabolit auch noch eine Rolle spielt)
- Metabolite stehen nicht mehr für andere Reaktionen zur Verfügung
- Zelle fliegt aus dem Fließgleichgewicht

TURBO DESIGN

schnell und trotzdem kontrolliert auf wechselnde Bedingungen zu reagieren gilt als grosse Herausforderung für eine Zelle!

- „**Turbo-Design**“: Pathways, die mit einem Schritt beginnen, der **ATP zur Aktivierung kostet**, später im Pathway aber **mehr ATP produzieren** als anfangs verbraucht wurde



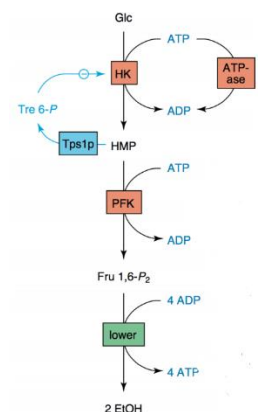
- **+) kleine Investition wird später durch eine grössere Ausbeute ausgeglichen**
- **-) bauen darauf, dass kontinuierliche Mengen an Substrat zur Verfügung stehen**
- **-) Substrate-Accelerated Death**
- wechselnde Einflüsse können unterschiedliche Verhaltensweisen von Pathways auslösen → für Pathways mit „Turbo -Design“ heisst das, dass sich unter plötzlich sehr hohen Substratkonzentrationen ein **anderer steady-state** einstellen kann als unter kontinuierlichem Substrateinfluss, **was schädlich für die Zelle** sein kann
- Rückkopplungsmechanismen helfen dabei, einen **bestimmten Gleichgewichtszustand in der Zelle beizubehalten** und ihr so **Robustheit gegenüber wechselnden Bedingungen** zu geben

SUBSTRATE-ACCELERATED DEATH

Eine Zelle stirbt daran, dass ihr zu viel eines bestimmten Nährstoffs zur Verfügung steht

BEISPIEL:

- Glykolyse als «Turbo-Design»
- Genprodukt von **Tps1** hemmt den ersten Schritt der Glykolyse
- Hefestamm, bei dem das Gen **Tps1** deletiert ist (tps1-Mutante) zeigt **substrateaccelerated death** durch das Substrat Glucose → „Turbo-Design“ der Glykolyse dazu, dass sie an plötzlichem Glukoseüberschuss sterben!



steady-state für die ATP -Konzentration: ATP-Produktion im 3. Schritt der Glykolyse =! Verbrauch von ATP in den ersten beiden Glykolyse-schritten und ATPase -Reaktion

Es gibt **zwei mögliche steady-states**:

FALL 1: kontinuierliche Mengen an Glucose:

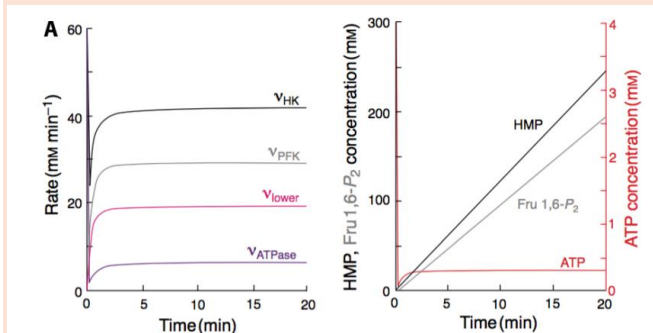
ATP für die Aktivierung der Glukose und relativ viel ATP für die ATPase-Reaktionen → Reaktionsgeschwindigkeiten sind in allen drei Schritten der Glykolyse gleich gross und die Geschwindigkeit der ATPase -Reaktion doppelt so hoch

FALL 2: Es gibt sehr viel Glucose:

Die ersten beiden Schritte der Glycolyse verbrauchen den Grossteil des verfügbaren ATPs wobei nun für die ATPase -Reaktionen nur sehr wenig ATP zur Verfügung bleibt

→ **2 PROBLEME** EINES NICHTREGULIERTEN TURBO-DESIGNS

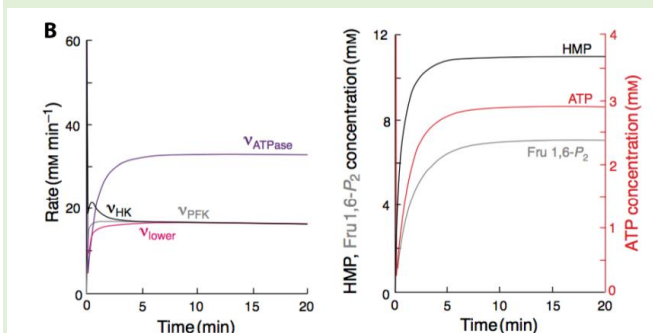
- **wenig ATP** für andere zellulären Prozesse zur Verfügung
→ Steady-state von ATP sinkt um mehr als das 6 Fache
→ Zelle stirbt an Energiemangel, obgleich sie Zucker hat
- **Konzentrationen akkumulieren** → Akkumulation von Intermediärprodukten der ersten beiden Reaktionen, welche toxisch sein können



DURCH RÜCKKOPPLUNGSMECHANISMEN REGULIERTER TURBO-DESIGN:

FALL 2 wird durch Rückkopplungsmechanismen verhindert!

→ Inhibition der Hexokinase durch Akkumulation von Hexosemonophosphat (Produkts des ersten Schrittes) → stellt sicher, dass sich ein steady-state für die ATP -Konzentration einstellt, bei welchem **keine Zwischenprodukte akkumulieren** und der **ATP-Spiegel in der Zelle hoch bleibt** → verhindert substrate-accelerated-death durch zu hohe [Glucose]



- Zelle befindet sich im Fließgleichgewicht → nichts akkumuliert, ATP Konzentration ist hoch genug

W4: REGULATION VON STOFFWECHSELWEGEN

Regulationsmechanismen werden benötigt, damit die Zelle angemessen und schnell auf sich wechselnde Bedingungen anpassen kann

REPETITION:

Warum brachen Zellen solche Reaktionsmechanismen?

Warum sollten Stoffwechselwege reguliert werden?

- Verhinderung starker Antworten
- Um schnell wieder ein steady-state zu erreichen
- Um flexibel auf Umweltveränderungen reagieren zu können → an ändernde Umweltbedingungen anpassen
- Um Robust gegenüber diversen Störungen zu bleiben → bei kleinen Schwankungen nicht zu schnell ändern
- Vermeiden eines Overshoots der Konzentrationen → Ressourcenverschwendung, Toxizität

METABOLISCHEN FLÜSSE UND METABOLITKONZENTRATIONEN

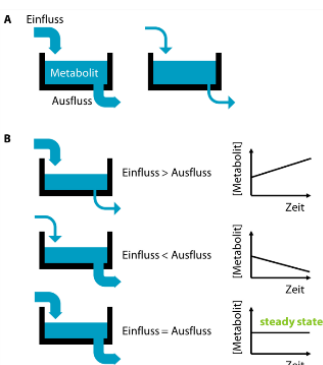
metabolischer Fluss = Geschwindigkeitsrate v einer Stoffwechselreaktion (→ Reaktionsrate)

- reversible Reaktion im thermodynamischen Gleichgewicht → Fluss = 0
- steady state → Fluss kann grösser als 0 sein

WECHSEL DER UMGEBUNG:

Der Metabolismus reagiert darauf, indem sich **die Flüsse durch die verschiedenen Stoffwechselreaktionen verändern**

→ Dies bedeutet aber **nicht** unbedingt, dass die Konzentrationen der Metaboliten verändert werden!



→ Flüsse beeinflussen die Veränderung der Konzentration, **hängen aber nicht direkt mit der absoluten Konzentration zusammen** (absolute Metabolitenkonz. sagt nicht über die Grösse der Flüsse aus!)

Metabolitenkonzentrationen und metabolische Flüsse stehen in folgendem Zusammenhang:

- Die Differenz aus Einfluss und Ausfluss bestimmt die **Konzentrationsveränderung** eines Metaboliten.
- Die Metabolitenkonzentrationen beeinflussen die metabolischen Flüsse durch die Abhängigkeit der **Reaktionsrate** von der Substratkonzentration

FAKTOREN DIE DIE REAKTIONSRATE BEEINFLUSSEN

Eine Enzymreaktion läuft schneller ab, wenn man:

1. **mehr Enzym** zur Verfügung hat ($[E]$)
2. die Enzyme **mehr Substratmoleküle pro Zeiteinheit umsetzen** können (k_{cat}) (spezifische Aktivität eines Enzyms wird verändert)
3. das Enzym eine **höhere Affinität** für Substrat hat (K_m)
4. die **Substratkonzentration steigt** ($[S]$)
5. Thermodynamik (Temperaturveränderung, Aktivierungsenergie, Fließgeschwindigkeit etc)

$$v = [E]k_{cat} \frac{[S]}{K_m + [S]}$$

Which physico-(bio)chemical processes contribute to the emergence of a new metabolic flux pattern?

- **Reaktionskinetik**
- **Thermodynamik**
- **Enzym Anwesenheit**
- **Enzym Aktivität**

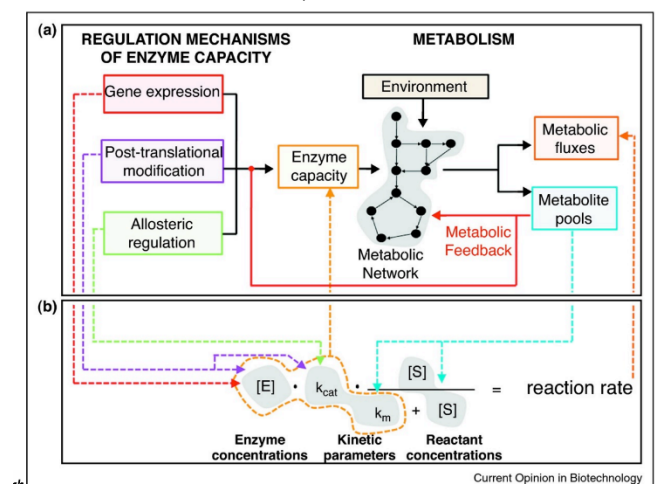
→ Nur die letzten 2 Punkte kann die Zelle aktiv verändern

→ Möchte ich mehr Fluss, brauche ich mehr Enzyme (v_{max} wird nicht verändert, nur $[E]$ und dadurch auch die Anzahl umgewandelter Substrate)

VERÄNDERUNG DER FAKTOREN DURCH REGULATIONSMECHANISMEN

Regulationsmechanismen, welche die Faktoren verändern, welche die Reaktionsrate beeinflussen:

- Veränderung der **Genexpression** → Je mehr ein Gen für ein Enzym exprimiert wird, desto grösser ist dessen Konzentration (beeinflusst $[E]$)
MERKE: Genexpression kann nur die $[E]$ verändern → nicht kinetische Faktoren!
- **Post-translationale Modifikationen** → Konzentrationsveränderung und Umsatzrate
MERKE: Posttranslationale Modifikationen können $[E]$, k_{cat} und K_m verändern (auch kinetische Parameter!)
- **allosterische Regulation** → Reguliert Umsatzrate (k_{cat})
MERKE: allosterische Regulation kann nur kinetische Parameter verändern, nicht die Enzymkonzentration
- **Substratkonzentration** → beeinflusst die Reaktionsrate
Abbau verstärken, Enzym-Ort verändern etc.



Der Einfluss der einzelnen Mechanismen auf die metabolischen Flüsse kann durch Trennung der Zeitskalen untersucht werden (Genexpression langsam, Modifikationen schnell)

MERKE:

- Schauen, ob Feedbackmechanismen stattfinden und ob sie positive/negative Wirkung haben (qualitativ)
- Stärke der Regulation betrachten (quantitativ)

ZIELSETZUNGEN

In einem Pathway findet man häufig **mehrere verschiedene Regulationssysteme** → **Verschiedene Zielsetzungen** bedingen unterschiedliche Regulationsmechanismen

Wie erreicht die eukaryotische Zelle trotz schwankender Nachfrage eine Energie-Homöostase (ATP)?

Feedback.

Aber warum so viele Feedbackschleifen?

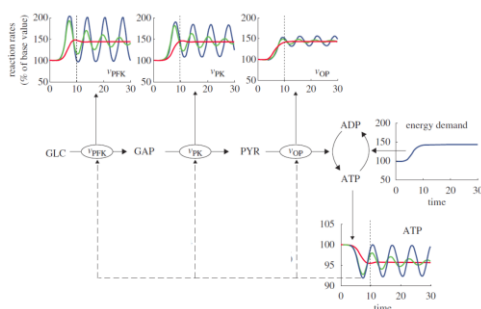
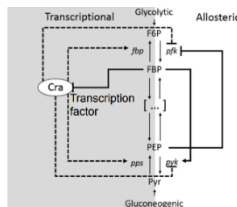
- Verschiedene Arten von (negativem) Feedback haben unterschiedliche Funktionen → Je nach Ziel werden andere negative Feedbacksysteme verwendet:
- Mit nur einem Feedback könnte das zu Oszillationen führen → durch **zeitliche Verschiebung**

Was ist das Ziel einer Regulation des Energie-Metabolismus?

- **ATP-Konzentration sollte fixiert sein** → auf demselben Level gehalten werden → **[ATP]=const.**
 - Es sollte ein **steady-state** geben (kein oszillierendes oder chaotisches Verhalten)
 - Regulation sollte dafür sorgen, dass bei einer Störung diese **möglichst schnell behoben** wird
- Je nach Ziel wirkt ein anderes Feedbacksystem

ZIEL 1) EINSTELLUNG EINER KONSTANTEN ATP-KONZENTRATION (PROPORTIONAL FEEDBACK)

- Negative Feedbacksysteme (Oft **allosterisch**), in jenen hohen ATP-Konzentrationen Enzyme inhibieren, findet man an **mehreren Stellen** im Energiemetabolismus
- Allosterische Regulation führt zu einer negativen Rückkopplung des Stoffwechsels **in Abhängigkeit von der aktuellen ATP-Konzentration**.
- **Phosphofructokinase (PFK)**, **Pyruvatkinase (PK)** etc.
- o PEP hemmt pfk (Hemmung nach oben)
- o F6P aktiviert pyk (Aktivierung nach unten)
- Falls nur ein Feedbackmechanismus vorhanden ist und alle anderen Feedbacksysteme inhibiert werden beginnen die [ATP] und die Reaktionsrate zu oszillieren



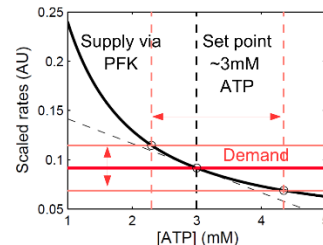
- Test von Einzelfeedbacks durch in silico Knock-Outs
- **Nur ein Feedbackmechanismus ist aktiv (Nur 'langes' PFK Feedback → Zeitverzögerung → Instabilität!)**
- Problem: zwischen der Reaktion und dem Feedbacksignal besteht eine **zeitliche Verzögerung** (Akkumulationen) → Lösung: mehrere FBS!
- **Zwei FBM sind aktiv**
- **Alle FBM sind wieder aktiv**

Durch **mehrere, nacheinander geschaltete Feedbacksysteme** stellt sich eine **konstante ATP-Konzentration** ein → verhindert Instabilität und ungewollte Verhaltensweisen wie Oszillation

PROBLEM: Was ist nun falsch an der ATP Antwort?

→ geht nicht auf 100% zurück!

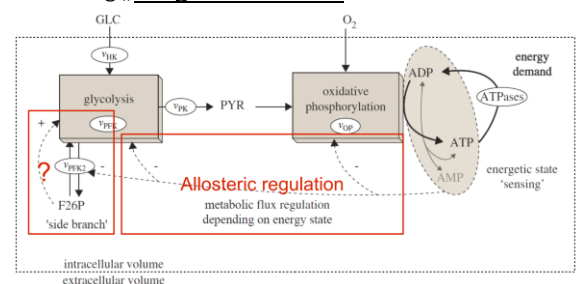
Proportionales Feedback führt für unterschiedliche **Demands** zu unterschiedlichen steady-state ATP-Konzentration → Grosse Abweichungen vom Sollwert (Erhöhter Verbrauch → Schnittpunkt wird zu niedrigeren ATP Konzentrationen verschoben → zu tiefer steady state)



Lösung: Weiteres negatives Feedback durch AMP:

ZIEL 2) EINE FIXE ATP KONZENTRATION, DIE UNABHÄNGIG VOM ATP-BEDARF IST

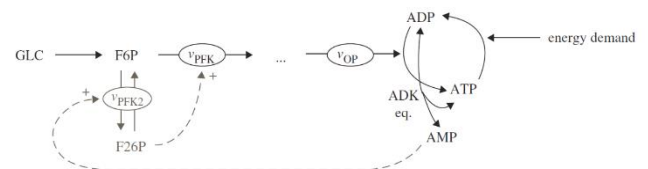
- **PROBLEM:** Durch Ansteigen des ATP-Bedarfs würde sich mit nur proportional feedback Regulation **eine gesenkte steady-state-Konzentration** einstellen
- Lösung „**integrales Feedback**“



→ weiteres Regulationssystem/ Feedback loop:

AMP-Konzentration als Indikator für die Abweichung der ATP-Konzentration von der vorherigen steady-state-Konzentration

- Ansteigen der AMP-Konzentration **aktiviert die F26P Produktion**, welches wiederum **PFK aktiviert** → Aktivierung der Glykolyse (F26P als eines der stärksten Aktivatoren der Glykolyse)



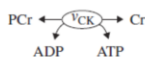
Seitenreaktion: **Phosphofructokinase 2 (PFK2)**

- Feedback ist anders: **Integrale** Regelung ist **nicht proportional zum aktuellen Signal**, sondern speichert die Abweichung **über die Zeit hinweg** → Betrachtet Veränderungen über die Zeit und auch nur kleinste Abweichungen führen dazu, dass PFK stark reguliert wird
- Ohne integrales Feedback: Abweichung vom ATP-Sollwert

ZIEL 3): AUFRECHTERHALTUNG DES ATP-SPIEGELS UNTER SCHNELLEN NÄHRSTOFFÄNDERUNGEN (DIFFERENTIELLES FEEDBACK)

- **Kreatinkinase:** zusätzliches Enzym, das in nährstoff-armen Phasen die Aufrechterhaltung der ATP-Konzentration sicherstellt (**Phosphokreatin als Puffer**)
→ überträgt bei hoher ATP Konzentration Phosphatrest von ATP auf Kreatin
→ überträgt bei tiefem ATP-Spiegel Phosphatreste von Phosphokreatin zurück aufs ADP (**Puffersystem**)
- Hat Hefe nicht → Bei Glukoseentzug fällt [ATP] unter Normal-Konzentration von 3mM
- **Pufferfunktion:**
- Puffer reagieren auf die Änderungsrate (differentiell) der Störung → kurzfristiger, schneller Störung wird effizienter entgegengewirkt

Dynamics of Creatine Kinase



- Mass-action kinetics for creatine kinase (CK):

$$v_{CK} = k_f[PCr][ADP] - k_r[Cr][ATP]$$

- With a small perturbation in ATP concentration, Δ :

$$v_{CK} = k_f[PCr]([ADP] + \Delta) - k_r[Cr]([ATP] - \Delta)$$

- Re-arranging the terms:

$$v_{CK} = (k_f[PCr][ADP] - k_r[Cr][ATP]) + (k_f[PCr] + k_r[Cr])\Delta$$

- Depending on parameters:

$$v_{CK} \propto -\frac{d[ATP]}{dt}$$

OMICS DATA

= Messungen von allen Komponenten aller Metaboliten, mRNAs oder Proteinen innerhalb einer Zelle

- Metabolomics: Messung aller Metabolitenkonzentrationen in einer Zelle, Körperflüssigkeit etc.
- Konzentrationen messen (durch Massenspektroskopie oder NMR → mit dem Ziel, möglichst viele Metabolitenkonzentrationen zu messen)
- Konzentrationen und Konzentrationsveränderungen an verschiedenen Stellen messen
- Schwierig genau zu sagen, was ist eine relevante Veränderung und was ist zufällig → Wir messen zu viel, als wir überhaupt interpretieren/ auswerten können

TURNOVER

Turnover Zeit eines Metabolitenpools hängt von der «Pool»-Grösse UND vom Durchfluss durch ihn hindurch ab

FLUSS

Beziehung zwischen Metabolit[conc] und Flüssen sind **Massenbilanzen** und **kinetische Gleichungen**.

Veränderungen der Metabolitenkonzentration ist extrem schnell → Dynamik ist extrem schnell

Modell um die Metabolitkonzentration als Flussveränderung interpretieren zu können:

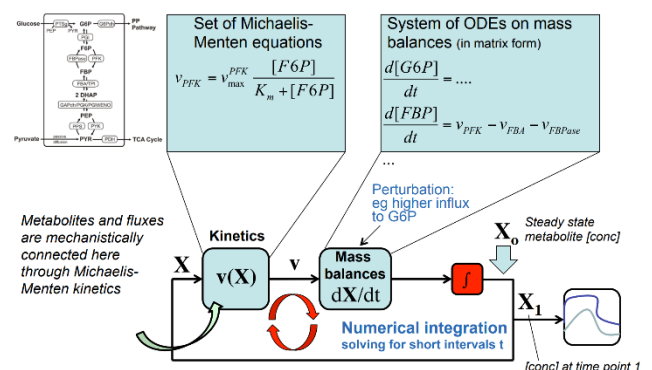
Fluss: Mittels Michaelis-Menten Kinetik kann die Reaktionsgeschwindigkeit berechnet werden

ABER: Fluss ist auch durch Konzentrationsveränderung der vorherigen Substraten ab → Zusammenspiel mit den anderen Flüssen betrachten!! Man muss eine Massenbilanz durchführen → ODEs verwenden!

Ohne Kinetik kann man aber nichts über das Verhalten der Konzentrationsauswirkungen auf die Flüsse sagen

Man braucht die beide Konzepte:

- Bilanz der Flüsse
- Metaboliten Flüsse via Michaelis-Menten



W5/6 ANALYSE METABOLISCHER NETZWERKE

MODELLE GENOMWEITER STOFFWECHSELNETZWERKE

- sind **stöchiometrische** Modelle
- sind **statische** Modelle (nicht dynamisch!)

Warum nicht dynamisch?

Zu komplex! Bräuchte zu viele Informationen um ein dynamisches Modell des gesamten Stoffwechsels zu entwerfen:

- Struktur des Netzwerks
- Für jede Reaktion bräuchte man Angaben über:
 - o Die initialen Metabolitkonzentrationen
 - o die kinetischen Parameter jedes Enzym (Enzymkonzentration, K_m und K_{cat})
 - o die Art der Reaktionskinetik (z.B. Mass-Action-Kinetik, Michaelis-Menten-Kinetik oder Substratinhibition)
 - o sämtliche regulatorischen Mechanismen

→ 2 Probleme:

- **Informationen** sind bei weitem noch **nicht** für alle zellulären Reaktionen eines Organismus **verfügbar**
- dynamische Simulation eines so grossen Netzwerks erfordert eine **enorme Rechenleistung**

ANNAHME:

Alle **Metabolitkonzentrationen** seien im **steady state**

im steady-state können wir **allein mit Massenbilanzen** arbeiten, weil alle Flüsse ausgeglichen sind und die Konzentrationen konstant bleiben

→ VEREINFACHUNG

- Die Grundannahme von Modellen genomweiter Stoffwechselnetzwerke ist der **steady state**.
- Bekannt ist dabei die Stöchiometrie aller Reaktionen.
- Gesucht sind die Werte der metabolischen Flüsse unter bestimmten Bedingungen

STOFFWECHSEL IM STEADY STATE

steady state = in einem Pool (state) ist die Summe der Zuflüsse gleich der Summe der Abflüsse

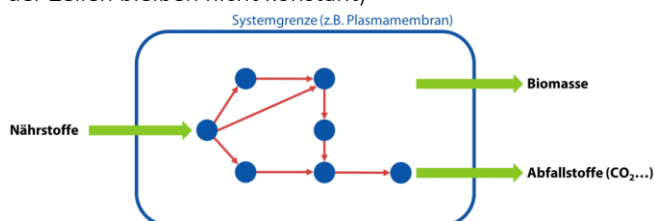
Im **Stoffwechsel**: jeder einzelne Pool der zellinternen Metabolite muss diese Bedingung erfüllen!

→ Alle Reaktionen innerhalb der Zelle laufen mit konstanter Geschwindigkeit ab

→ Einfluss = Ausfluss → **Zellinterne Flüsse** sind balanciert

→ **Konzentrationen** bleiben konstant

→ **Transportraten** in die Zelle und aus der Zelle sind konstant (**Konzentration extrazellulärer Metaboliten** und die Anzahl der Zellen bleiben nicht konstant)



MODELL-VERGLEICH

BISHERIGE MODELLE: DYNAMISCH - ODEs

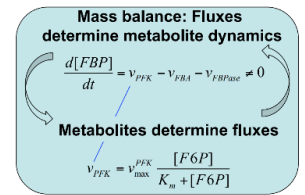
- können **kleine Pathways** anschauen
- Sowohl Mass-Balance als auch Kinetische Modelle
- Parameter Identifikation

So kann man **kleine Pathways**

anschauen, jedoch braucht man

dazu **alle Parameter** für jedes

Enzym → geht nicht für gesamte Stoffwechselwege



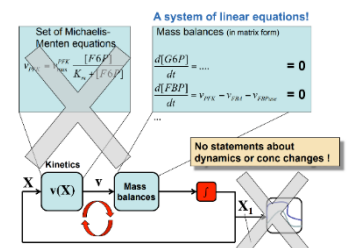
MODELLE FÜR GESAMTE STOFFWECHSELWEGE

STATISCH - FBA

- Einschränkung/Annahme: **steady state!**
→ simplification to **linear mass balance equations**

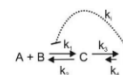
- Kein Gebrauch kinetischen Modelle mehr, keine Parameter nötig

- Man verwendet **nur noch** «linear mass balance equations», beachtet die Kinetik also nichtmehr!



MEKRE: Ich kann nun aber keine Annahmen mehr über Konzentrationsveränderungen, sondern nur noch über die steady-state Situation machen!

Model class	Level of abstraction	Required information	Example applications
Topological (steady state)	Interaction	Components and unspecified connections must be known	- Genetic networks - Protein-protein interaction - Metabolite-protein interaction
Steady state	Reaction stoichiometry	Mass and energy balances thermodynamics (directionality)	Metabolic networks - flux balance analysis - elementary flux modes - ...
Mechanistic (dynamic)	Enzyme mechanism and regulation	Kinetic parameters	Kinetic models (including regulation)



METABOLISCHE FLUSSMUSTER

Gesucht: wie gross sind die Raten der verschiedenen metabolischen Reaktionen unter bestimmten Bedingungen?
→ Constraint-based modeling → FBA

Merke: es kann mehrere Lösungen für das Flussmuster eines grossen Stoffwechselnetzwerks geben, die trotzdem alle die Bedingungen erfüllen, dass sämtliche Metabolitkonzentrationen im steady state sind

Das optimale Flussmuster kann jedoch durch geeignete Einschränkungen und das Setzen einer Zielfunktion unter bestimmten Bedingungen ermittelt werden

CONSTRAINT-BASED STOICHIOMETRIC MODELING

= einschränkungsbasiertes Modellieren

Man bestimmt den Output des Modells (hier: Flussmuster des Reaktionsnetzwerks) dadurch, dass immer mehr

Einschränkungen getroffen werden:

- Für **Metabolismus**-Modeling geeignet
- Basierend auf mass balances (keine Parameter!)
- Nur **Stöchiometrie**
- Undeterminiert → es gibt keine perfekten Lösungen, sondern mehrere mögliche Wege!
- **Nur für steady-state** anwendbar!! (innerhalb!)
- **steady-state** = Fließgleichgewicht!
- es gilt also:* $\frac{dX}{dt} = v_1 + v_2 = 0$
- (v1= Einfluss, v2= Ausfluss)

WAS MAN KENNEN MUSS: (für Rekonstruktion von biologischem Stoffwechselweg)

- alle relevanten Reaktionen
- Reaktionsstöchiometrie und Richtungsabhängigkeit
- Cofaktoren & Cofaktor-Spezifität
- Lokalisierung von Reaktionen (Kompartimente)
- Spezielle Reaktionen (z.B. Atmungsketten)

FLUX BALANCE ANALYSIS

- typisches Beispiel für constraint-based modeling
- **Annahme:** Steady-state = Fließgleichgewicht
- quasie steady-state durch Zeit-skalen Separationen
- **Nichtdeterminiert**
- Wir haben ein nicht-determiniertes Gleichungssystem mit mehreren möglichen Lösungen
- Erreichen durch **Massenbilanzen** aller Metaboliten
- **Interesse:** Wie können wir den Lösungsraum einschränken und auf unser bestimmtes, biologisches System anwenden → Welchen Weg wählt die Zelle?
- Methode, mit der man das **optimale Flussmuster eines Stoffwechselnetzwerks** unter gegebenen Bedingungen berechnen kann

Es gibt viele verschiedene Möglichkeiten, welche Raten die Stoffwechselreaktionen im steady state haben können.

Durch Flux Balance Analysis bestimmt man, welches Flussmuster unter bestimmten Kriterien optimal ist.

WELCHE INFORMATIONEN BRACHT MAN?

- Komponenten und Vernetzungen
- Reaktionsstöchiometrie
- Massen und Energiebalancen
- Thermodynamik
- o Stöchiometrische Matrix
- o Constraints (steady state, (IR)Reversibilität)
- o Zielfunktion → Objectives

VORGEHEN FLUX BALANCE ANALYSIS

1. Setzen möglichst vieler **Einschränkungen**
2. Wahl einer **geeigneten Zielfunktion**
3. Bestimmung jenes Flussmusters, welches die Zielfunktion möglichst gut erfüllt

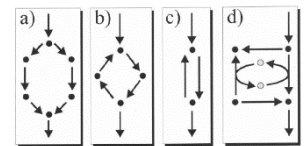
Flux Balance Analysis ist ein Beispiel für constraint-based modeling: Man **schränkt den Lösungsraum immer weiter ein** und nähert sich so der Lösung an. Um ein eindeutiges Flussmuster zu erhalten, muss man eine geeignete **Zielfunktion** wählen

Auf der Suche nach den Reaktionsraten schränkt man den Lösungsraum möglichst weit ein

PITFALL: FBA-Vorhersagen hängen stark von der Wahl der Zielfunktion ab (z.B. Wachstum als Ziel) → Nicht immer (jeder Organismus, alle Bedingungen) zutreffend

SITUATIONEN, IN WELCHEN STÖCHIOMETRISCHE FLUX ANALYSIS SCHEITERT (UNDETERMINIERT):

- (a) parallele Pathways ohne related flux measurement
- (b) metabolische Zyklen (futile)
- (c) bidirektionale Reaktionen
- (d) aufgeteilte Pathways bei welchen die Cofaktoren (grau) nicht balanciert sind



Umgehen:

- Constraints hinzufügen → Solutionspace minimieren
- Datensätze hinzufügen (z.B. isotopische Tracer)
- Objective functions hinzufügen

Um ein undeterminiertes System einzuschränken, kann man

- **«Constraints»** hinzufügen (how do we know that a reaction actually could take place?)
- Starke, biologische Annahmen treffen (e.g. on ATP production, cofactor balancing or evolutionary objectives) → **«Objectives»**

→ Es gibt viele Situationen, wo man keine klare Lösungsmenge erhält und dadurch mittels Constraints den Lösungsraum immer mehr einschränken und schlussendlich durch Objectives den optimalen Weg finden kann

MÖGLICHE EINSCHRÄNKUNGEN:

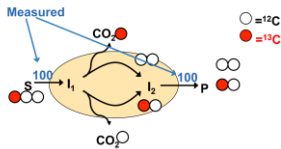
1. **Struktur des Netzwerks** (festlegen, welche Interaktionen zw. Metaboliten und Reaktionen es überhaupt gibt)
2. **steady-state-Annahme** (Einschränkung des Lösungsraum und brauchen keine Angaben über dynamische Prozesse mehr)
- **(Ir)reversibilität?** (Flüsse irreversibler Reaktionen sind nicht kleiner als 0)
- **Obergrenzen** der Reaktionsraten
- Informationen über **Bedingungen**, in welchen sich die Zelle befindet
- **Transportrate** von Austauschflüssen (über die Zellmembran einfach messbar)

CONSTRAINTS:

Schafft zusätzliche Informationen

Einschränkung des Lösungsraumes durch:

- **Principle-based Constraints:** Prinzipienbasiert (Reaktionsdirektionalität → Thermodynamik etc.)
 - **Data-based Constraints:** (Isotopische Markierer etc.) → NMR oder MS-Analyse
- ACHTUNG: ATP Konzentration dient nicht als Indikator! (ATP-Konzentration in der Zelle wird konstant gehalten!)



MERKE: Constraints schränken den Lösungsraum null space
 → feasible space) nur ein, finden aber meist noch keine
 optimale Lösung

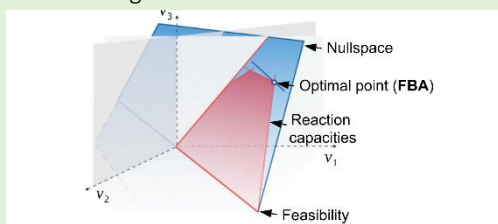
OBJECTIVE FUNCTIONS:

Biologisch mögliche objective Functions:

→ Optimierung/ Eindeutiges Flussmuster durch **Annahme**
 einer Zielfunktion

- baut darauf, dass sich biologische Systeme durch die Evolution so entwickelt haben, **dass sie bestimmte Ziele möglichst gut verfolgen**, da Organismen in ständiger Konkurrenz mit anderen Lebewesen stehen
- Zielfunktion drückt als Wert aus, **wie gut das Stoffwechselnetzwerk ein bestimmtes Ziel erfüllt**
 - o Maximierung des Wachstums
 - o ATP/ Energieformierung
 - o Bevorzugung eines direkten Weges → möglichst wenig Flüsse/ Wege → Gesamtverkehrsfluss minimieren

- **Constraints:** Null space soll auf den Feasible space eingeschränkt werden → Man sammelt Daten und Annahmen und schränkt dabei den Lösungsraum ein
- **Objectives:** Aus dem feasible Space (möglichen) soll nun den optimalen Weg gefunden werden → Man überlegt sich das Ziel der Zelle und wählt dadurch den objektiv optimalen Weg aus



□ **Constraints** define the search space ('valid' vs. 'invalid').

□ **Objectives** encode biological assumptions on optimality.

Constraints

□ Quasi steady state assumption: $0 = S \cdot v$

□ Reaction reversibilities / capacities: $\alpha_i \leq v_i \leq \beta_i$

□ Optimal feasible steady state: $w^T \cdot v \rightarrow \max!$

Objectives

□ The weights w encode the assumed cellular objective function, **here** maximal growth rate:

$$w_{\text{growth}} = 1, \quad w_{\text{all other fluxes}} = 0$$

Constraints und Objectives erlauben es, Regelungs-, Störungs- und Omics-Daten in stöchiometrische Modelle mit entsprechenden (modifizierten) Algorithmen einzubeziehen

PSEUDOCODE FÜR EINE FLUX VARIABILITY ANALYSIS

Inputs: Stoichiometric matrix and upper / lower bounds for the fluxes

Output: Minimal and maximal fluxes

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

Inputs: Stoichiometric matrix S ($m \times n$)

Flux bound vectors α and β ($n \times 1$)

Output: Matrix of minimal and maximal fluxes F ($n \times 2$)

Set F to $n \times 2$ zero matrix // initialization

For $d = 1$ to 2 // loop over the two reaction directions

For $z = 1$ to n // loop over all reactions

Set w to $n \times 1$ zero vector

Set w_z to $(-1)^d$

Solve linear program $w^T \cdot v \rightarrow \max!$
 $s.t.$

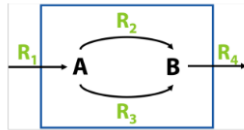
$$S \cdot v = 0$$

$$\alpha_i \leq v_i \leq \beta_i$$

Set $F_{z,d}$ to v_z

VORGEHEN BEI DER FLUSSMUSTER SUCHE

- vier Reaktionen (R1, R2, R3, R4)
- zwei Metabolite (A und B)



- **Bedingung: Konzentrationen von A und B müssen gleichbleiben (steady state!)**
- Ziel: Wie gross sind die Reaktionsraten der vier Reaktionen (v_1, v_2, v_3, v_4)?

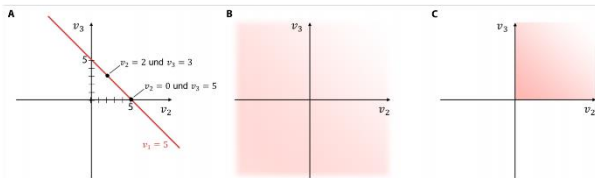
1. Aufstellen der Massenbilanzen für beide Metabolite

$$\frac{d[A]}{dt} = v_1 - v_2 - v_3 = 0 \quad \text{(steady-state-Bedingung)}$$

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

2. Es gilt: $v_1 = v_4$

- Wir haben nun aber **mehr unbekannte Variablen als Gleichungen (mehr Reaktionen als Metabolite)**
→ KEINE EINDEUTIGE LÖSUNG MÖGLICH!
- verschiedenen Lösungsmöglichkeiten bilden den **Lösungsraum** (engl. solution space)



- Falls v_1 gegeben ist (=5) liegen die Werte für v_2 & v_3 auf einer Geraden
 - v_1 sei kann beliebig gross sein → Lösungsraum unendlich grosse Fläche
 - Irreversible Flüsse können nicht kleiner als 0 sein** (Thermodynamische Informationen sind sehr nützlich für die Einschränkung des Lösungsraums) → Lösungsraum nur der 1. Quadrant
- **Irreversible Reaktionen** (nur einen Pfeil) können nur **positive Raten** haben → Lösungsraum befindet sich immer nur im **1. Quadrant**

STÖCHIOMETRISCHE MATRIX (S)

=mathematische Repräsentation eines Reaktionsnetzwerks

- **Spalten:** repräsentieren die **Reaktionen** des Netzwerks
- **Zeilen:** repräsentieren die **Metaboliten**

$$S = \begin{pmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \end{matrix}$$

- **1 (erste Zeile & ersten Spalte):** Metabolit A entsteht in der Reaktion R1 einmal pro Reaktion
- **-1 (erste Zeile der zweiten Spalte):** Metabolit A wird in der Reaktion R2 einmal pro Reaktion verbraucht

Raten der vier Reaktionen im steady state: $\frac{dc}{dt} = S \cdot v = 0$

- $\frac{dc}{dt}$ = Konzentrationsveränderungen aller Metaboliten
- Gesucht: **Vektor der Reaktionsraten** $v = (v_1, v_2, v_3, v_4)$

$$S \cdot v = \begin{pmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 1 \cdot v_1 + (-1) \cdot v_2 + (-1) \cdot v_3 + 0 \cdot v_4 \\ 0 \cdot v_1 + 1 \cdot v_2 + 1 \cdot v_3 + (-1) \cdot v_4 \end{pmatrix}$$

- erste Zeile des Ergebnisvektors:

Konzentrationsveränderung des Metaboliten A

- zweite Zeile: Konz.veränderung für den Metaboliten B

$$\frac{d[A]}{dt} = 1 \cdot v_1 + (-1) \cdot v_2 + (-1) \cdot v_3 + 0 \cdot v_4$$

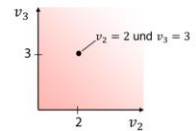
$$\frac{d[B]}{dt} = 0 \cdot v_1 + 1 \cdot v_2 + 1 \cdot v_3 + (-1) \cdot v_4$$

KERNEL-MATRIX

Enthält **linear unabhängige Lösungen des Gleichungssystems**

- linear unabhängigen Lösungen des Gleichungssystems: Vektoren, aus denen man durch Linearkombination alle anderen Lösungen erreichen kann
- Linear unabhängig Vektoren bilden die **Kernel-Matrix**

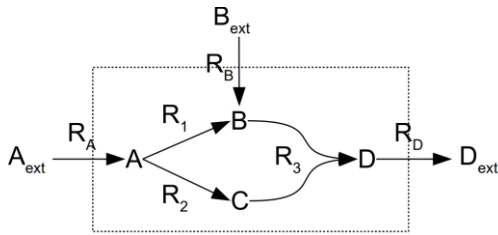
$$K = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{matrix} \leftarrow R_1 \\ \leftarrow R_2 \\ \leftarrow R_3 \\ \leftarrow R_4 \end{matrix} \Rightarrow v = \begin{pmatrix} 5 \\ 2 \\ 3 \\ 5 \end{pmatrix}$$



Der Lösungsraum enthält alle verschiedenen Lösungen für das Flussmuster unter den gesetzten Einschränkungen. Alle Lösungsmöglichkeiten können durch eine Linearkombination der in der Kernel-Matrix enthaltenen Lösungsvektoren gebildet werden

BEISPIELE STÖCHIOMETRISCHE MODELLE

GESUCHT: 6 **Metabolische Flüsse (v)** im steady-state in diesem Netzwerk



1. Netzwerkstruktur → Wissen über Metabolite und Reaktionen des metabolischen Netzwerkes

- 4 Metabolite (A, B, C, D)
- 6 metabolische Reaktionen (R_i)

2. Wissen über Irreversibilität

3. Aufstellen der Massen-Bilanz jeder Metabolite

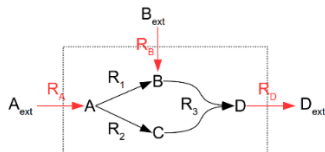
$$\frac{d[A]}{dt} = +v_A - v_1 - v_2 = 0$$

$$\frac{d[B]}{dt} = +v_B + v_1 - v_3 = 0$$

$$\frac{d[C]}{dt} = +v_2 - v_3 = 0$$

$$\frac{d[D]}{dt} = +v_3 - v_D = 0$$

4. Abhängige Flüsse erkennen: $v_2 = v_3 = v_D$



□ Reaction coupling allows for model simplification:

$$\begin{aligned} +v_A - v_1 - v_2 &= 0 \\ +v_B + v_1 - v_3 &= 0 \\ +v_2 - v_3 &= 0 \\ +v_3 - v_D &= 0 \\ \Rightarrow v_2 = v_3 = v_D \end{aligned}$$

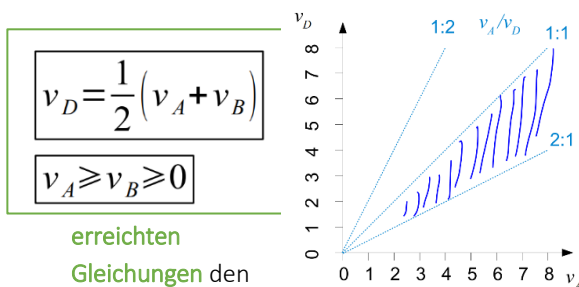
$$\begin{aligned} v_A &= v_1 + v_D \\ v_B &= v_D - v_1 \\ \text{Sum} &\Rightarrow v_D = \frac{1}{2}(v_A + v_B) \end{aligned}$$

5. Vereinfachungen treffen: Annahme der Irreversibilität (Flüsse müssen grösser als 0 sein!)

$$\begin{aligned} v_A &= v_1 + v_D \\ v_B &= v_D - v_1 \\ \text{Difference} &\rightarrow v_A - v_B = 2v_1 \\ \text{Sum} &\rightarrow v_D = \frac{1}{2}(v_A + v_B) \end{aligned}$$

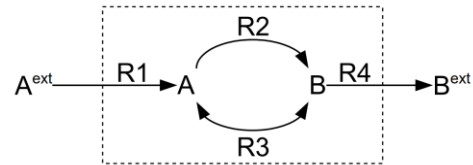
$$\Rightarrow v_D \geq \frac{1}{2}(v_A + v_B) \quad \Rightarrow v_A \geq v_B \geq 0$$

6. Mittels 2



erreichten
Gleichungen den
Lösungsraum
bestimmen

MATRIX ERSTELLEN



1. Netzwerkstruktur:

- 2 Metabolite (A, B)
- 4 metabolische Reaktionen (R1-4)

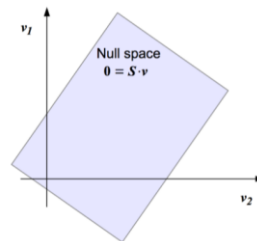
2. Stöchiometrische Matrix (S) bestimmen

$$S = \begin{bmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{bmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \end{matrix}$$

3. Quasi steady state → Metabolite balancing equation:

$$\frac{d c(t)}{dt} = S \cdot v(t) \xrightarrow{\text{const.}} 0 = S \cdot v$$

→ Solution Space: $0 = S \cdot v$

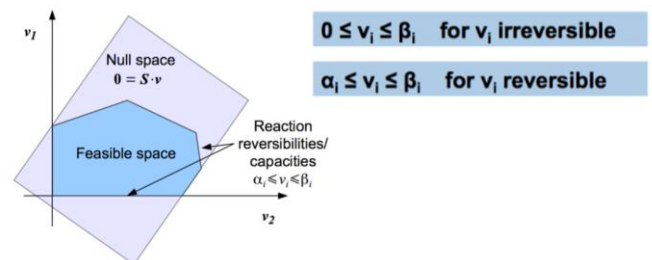


Immer eine Lösung: $v = 0 \rightarrow$ Thermodynamisches GGW

4. Feasible space: Weitere Einschränkungen durch Reversibilitäten und Kapazitäten (Constraints)

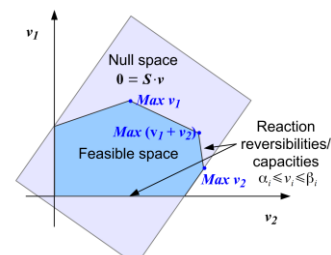
→ R3 ist **reversibel**, die anderen sind irreversibel

- Feasibility criterion: $v_i \geq 0$ for all irreversible reactions



5. Weitere Einschränkungen durch Zielfunktion Objectives) =Max. und Min. des Feasible Space

- Bei **Wachstum**: Optimal feasible steady state: where the weights w encode the assumed cellular objective function such as maximal growth rate
- Zellen wollen so schnell wie möglich wachsen.
- Die Zellen haben sich an dieses Ziel angepasst.



$$w^T \cdot v \rightarrow \max!$$

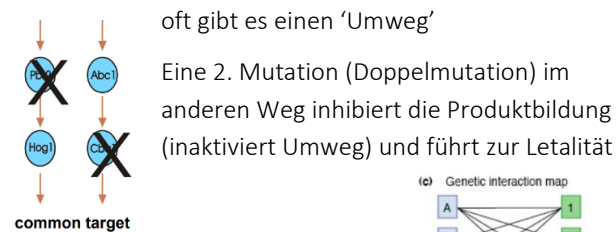
BEISPIELE ZU KOMBINATIONSTHERAPIE

- Komponente/ Reaktionen nicht als Einzelne betrachten, sondern als Netzwerke!
- Einzelmedikamente sind oft nicht ausreichend, da die Zelle die meisten Reaktionswege umgehen kann → Kombinationstherapie anwenden

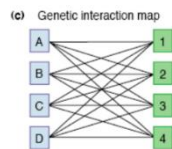
KREBSZELLEN

- Kollektion von genetischen Defekts
- Spezieller Stoffwechsel
- o Wachsen schnell
- o Warburg Effekt
- o Glutamin Abhängigkeit
- Entwickeln oft **Adaptionen gegen Medikamente**
- o Multidrugresistance (hinauspumpen)
- o Oberflächen/Struktur Veränderung
- o Umweg der Stoffwechselwege

DAS KONZEPT DER «SYNTHETISCHEN LETALITÄT» IN DER ENTWICKLUNG DER MEDIKAMENT-RESISTENZ



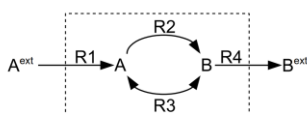
Durch die synthetische Letalität erkennt man Interaktionen



BEISPIEL ZELLWACHSTUM:

1. Um Zellwachstum (anabole Biochemie, Makromolekül Synthese) zu simulieren, brauche ich einen **Biomassen-Vektor**:
 - **Energie**, welche für Polymerisierung gebraucht wird
 - Wissen um **Makromolekül-Vorhandensein** in Zelle
 - **Zusammensetzung der Makromoleküle** einer Zelle
 - **Metabolite und Kofaktoren**
 - **Stöchiometrie** (Biomassezusammensetzung)
2. Einschränkungen treffen: Irreversibilität?
 - *Welche Metabolitenkonz. Bleibt konstant:?*
 - → *Steady-state-Annahme: innerhalb des Systems (Zelle) muss immer alles konstant bleiben! → E stellt den Kontakt zur Aussenwelt ist: Alles was ausserhalb meines Systems ist, kann sich verändern, nur der Fluss bleibt konstant!*

LETHALITÄT



□ Which **combinations of knockouts** are lethal?

- Single-knockouts of G1 and G4 are lethal.
- The combination of knockouts of G2 and G3 is lethal as well → **Synthetic lethality**.

§

FVA

FVA berechnet alle möglichen Flussverteilungen für alle Reaktionen mit allen Minimum und Maximum für eine bestimmte objective function.

ANTWORT AUF STÖRUNG: MOMA

- Annahme um Effekte der Gen Deletion vorherzusagen: Minimization of Metabolic adjustment (MOMA)
- Verwendet eine quadratische Zielfunktion

$$D(\mathbf{a}, \mathbf{c}) = \sum_{i=1}^n (a_i - c_i)^2$$

- MOMA zeigt den sofortigen Effekt eines Knock-Outs, während FBA den Effekt nach Erreichen des Optimums (→ Anpassung des Metabolismus) zeigt.
- Verschiedene Zielfunktionen können die Geometrie und die Rechenschwierigkeit der Optimierung verändern

MODELLIERUNG VON SIGNALTRANSDUKTIONSWEGEN

Die Zellsignalisierung folgt den gleichen Prinzipien wie für den Stoffwechsel: Bedeutung von Dynamik und Feedback.

Signaltransduktionsweg: Externes Signal wird empfangen, über eine Signalkette innerhalb der Zelle weitergegeben um dann eine zelluläre Antwort auszulösen

HERAUSVORDERUNG BEIM UNTERSUCHEN VON SIGNALTRANSDUKTIONSWEGEN: Beispiele:

- Crosstalk; Signal → Zellantwort
- Spezifität des Signals, Noise
- Regulation durch Feedbackmechanismen
- Dynamik, Parameter, Konzentrationen
- Amplifikation, Abschwächung
- Schalter als Schwellenwert → **Ultrasensitivität**

SENSITIVITÄT

BEGRIFFSDEFINITION

Sensitivität: Wie stark reagiert ein Signaltransduktionsweg auf Signale → wie stark verändert sich eine betrachtete Grösse, wenn sich ein bestimmter anderer Teil des Systems verändert wird

Sensitivität = Änderung des Werts einer Ausgangsgrösse im Verhältnis zur Änderung des Werts der Eingangsgrösse

$$\text{Sensitivität} = \frac{\text{Output}_2 - \text{Output}_1}{\text{Input}_2 - \text{Input}_1} = \frac{\Delta \text{Output}}{\Delta \text{Input}}$$

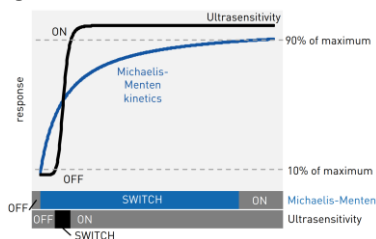
Beispiel: wie verändert sich die Aktivität von Kinasen (Output) abhängig von Änderung der Menge extrazellulär gebundener Signalmoleküle (Input)?

hohe Sensitivität führt zu „**schalterartigem**“ Verhalten in einem ganz bestimmten Bereich der Signalstärke → «Entweder ganz oder gar nicht»

Für biologische Systeme ist „schalterartiges“ Verhalten in einem bestimmten Konzentrationsbereich des Signals sehr **wertvoll** → erreichen einer kritischen Konzentration → Zellteilung, Nervenreizweiterleitung, Apoptose etc.

KURVENVERLAUF DER SENSITIVITÄT

Michaelis-Menten Kinetik: 81-fache Veränderung im Input wird gebraucht, um einen Wechsel von <10% (OFF) zu >90% (ON) zu ermöglichen



Ultrasensitivität: Ein System, welches **sensitiver reagiert** als durch eine Michaelis-Menten-Kinetik

- Starke Reaktion durch kleine Veränderung der Liganden Konzentration → Switch-Bereich wird stark verkleinert
- Sigmoidale Kurve
- Oberhalb des Switch-Bereiches gibt es aber Einbußen an Sensitivität
- Durch Ultrasensitivität kann ein „**schalterartiges**“ Verhalten erzielt werden, was für biologische Systeme sehr wertvoll ist, z.B. bei der Entscheidung zur Zellteilung.

HILL-GLEICHUNG:

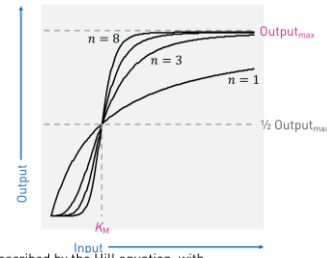
Die **Hill-Gleichung** beschreibt Ultrasensitivität quantitativ

→ beschreibt die **sigmoidale Kurve** der Ultrasensitivität

$$\text{Output} = \text{Output}_{\max} \frac{\text{Input}^n}{\text{Input}^n + \text{Input}_{0.5}^n}$$

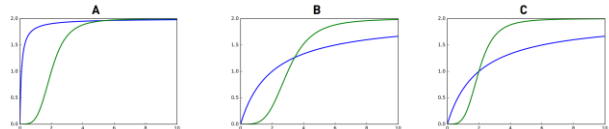
n: Hill-Koeffizient = Mass für die Ultrasensitivität: Je grösser n, desto steiler die sigmoidale Kurve und desto kleiner der Änderungsbereich beim Inputwert, der zu einer „schalterartigen“ Veränderung des Outputs führt
Input_{0.5}: Inputwert, bei dem der halbmaximale Outputwert erreicht ist

MERKE: Michaelis-Menten-Kinetik entspricht einer Hill-Gleichung mit dem Hill-Koeffizienten n=1



Draw two curves, both described by the Hill equation, with

- i) $\text{Output}_{\max}=2$, $K_M=2$, $n=4$ in green and
- ii) $\text{Output}_{\max}=2$, $K_M=2$, $n=1$ in blue



→ C ($K_M=2$) → müssen sich beim K_M (unten=2) schneiden!

WIE WIRD ULTRASENSITIVITÄT ERREICHT?

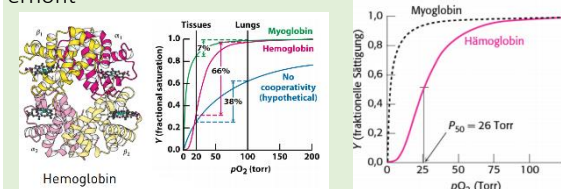
- Kooperativität
- Multistep Ultrasensitivität

1. KOOPERATIVITÄT:

Kooperativität der Sauerstoffbindung von Hämoglobin:

Hämoglobin und Myoglobin binden beide Sauerstoff, jedoch ist die **Bindungseffizienz von Hämoglobin** (Bluttransport) sehr viel höher als die von Myoglobin (Muskel als Sauerstoffspeicher)

Effizienzsteigerung des Hämoglobins wird erreicht durch seinen **Aufbau aus vier Polypeptidketten**, welche Sauerstoff „**kooperativ**“ binden → durch die Bindung eines O₂-Moleküls wird die Konformation der Nachbareinheit verändert, was die Affinität für das Binden eines weiteren Sauerstoffmoleküls erhöht



Hämoglobin: Sauerstoffbindekurve zeigt einen **sigmoidalen** (S-förmigen) Sättigungsverlauf → durch **Kooperativität** der vier Untereinheiten

Myoglobin: Michaelis-Menten-Kinetik → nicht-kooperative Reaktion

2. MULTISTEP ULTRASENSITIVITÄT

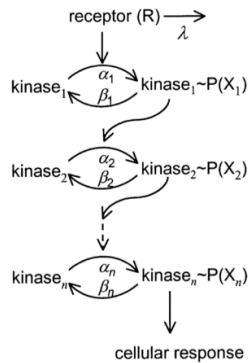


1 Effektor beeinflusst mehr als 1 Molekül aus dem Pathway

MODELLIERUNG VON PROTEINKINASEKASKADEN

VERALLGEMEINERTE SCHEMA EINER PHOSPHORYLIERUNGSKASKADE

- Aktive, phosphorylierte Kinase: X_i
- Inaktive Kinase: $\tilde{X}_i$
- Parameter für Kinase: α_i
- Parameter für Phosphatase: β_i
- Input: Rezeptor Signal $R(t)$
- R : aktivierter **Rezeptor** aktiviert die erste Kinase durch Phosphorylierung mit einer Geschwindigkeitskonstante α_1
- R wird mit Geschwindigkeitskonstante γ abgebaut



DIFFERENTIALGLEICHUNG:

Aktive Form der ersten Kinase der Kaskade X_1 :

$$\frac{dX_1}{dt} = v_{p,1} - v_{d,1} = \alpha_1 \tilde{X}_1 R - \beta_1 X_1$$

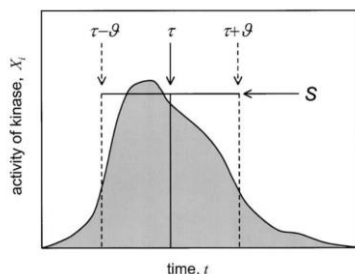
Alle weiteren Kinasen (Aktivität):

$$\frac{dX_i}{dt} = v_{p,i} - v_{d,i} = \alpha_i \tilde{X}_i X_{i-1} - \beta_i X_i = \alpha_i X_{i-1} \left(1 - \frac{X_i}{C_i}\right) - \beta_i X_i$$

Schlüsselfragen:

Welche Komponenten beeinflussen das Signal am meisten → welche sind wirklich wichtig?

- wie schnell wird ein Signal übertragen und die Zellantwort ausgelöst? (signaling time)
- wie lange dauert das in der Zelle ausgelöste Signal an? (signal duration)
- wie stark ist das Signal? (signal amplitude)



Analytische Lösung für die Leistung einer Kaskade:

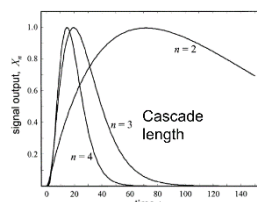
Signaling time τ : Expected time of arrival of the signal
 → hängt davon ab, wie die **Phosphatasen reagieren (β)**
 → α ist nicht relevant → Aktivität der Kinasen hat keinen Einfluss → **Kinaseindependent** (no α)!

Signal duration θ : Variance of the expected time
 → Phosphatasen haben einen Einfluss,
Kinaseindependent (no α)!

Signal amplitude S : Average amplitude during signaling
 → **Kinasen und Phosphatasen haben einen Einfluss** → **Kinasedependent!**

Qualitative Dynamik: Verstärkung oder Dämpfung des die Kaskade durchquerenden Signals ist **Parameterabhängig**

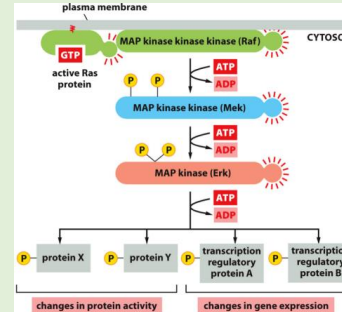
Gegen-Intuition: Kaskaden mit mehreren Levels können Signale schneller ausbreiten.



MAP-KINASE

MAP-Kinase-Signalweg: zeigt Ultrasensitivität

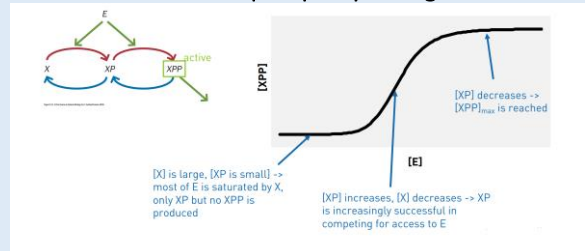
- Mitogen-activated protein kinase Signalweg:
- Drei Ebene von Proteinkinasen (hoch konserviert)
- Kontrolliert viele Prozesse (Wachstum, Differenzierung, Apoptose etc.)
- 1. Stufe (MAPKKK) wird oft von G-Protein Ras aktiviert
- 3. Stufe modifiziert dann andere Proteine und Transkriptionsfaktoren im Nucleus



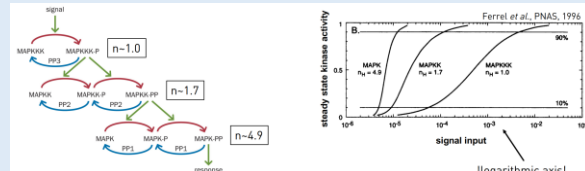
- MAPKK und MAPK müssen an **jeweils zwei Positionen phosphoryliert werden**, um die volle Aktivität zu erreichen

HOHE ULTRASENSITIVITÄT DURCH:

- Ultrasensitivität Antwort der MAPKK und MAPK basieren auf dem **Dualphosphorylierungs-Mechanismus**



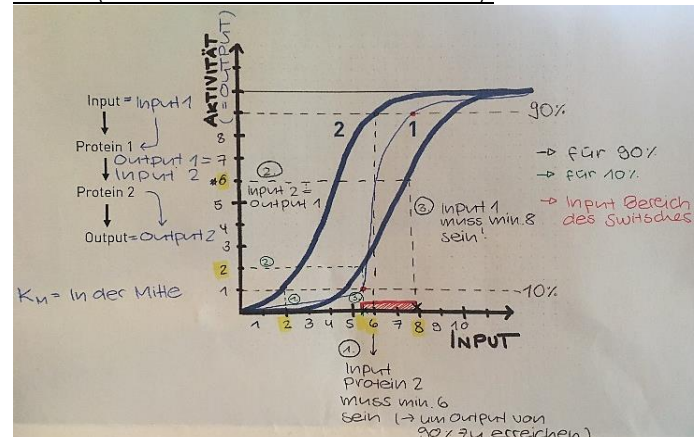
- MAPK-Signalweg erreicht seine hohe Ultrasensitivität durch **Kombination mehrerer ultrasensitiver Schritte**



→ Kombination führt zu einem noch höheren Hill-Koeffizienten! **Wieso?**

Input Kaskade= Input Protein 1, Output Protein 1 = Input Protein 2 und Output Protein 2 = Output Kaskade

Bestimme den Bereich des Inputs, bei welchem die Kaskade switcht (von 10% Aktivität auf 90% Aktivität):



SRC-TYROSINKINASE

Wie können manche Signalmoleküle je nach Signal unterschiedlich reagieren und so verschiedene Zellantworten bewirken?

Signalmoleküle, die an vielen Prozessen beteiligt sind, müssen:

- In ihrer Aktivität streng reguliert sein
- Oft verschiedene Konformationen einnehmen können

SRC: «Ein Signalmolekül mit vielfältigen dynamischen Verhaltensweisen» - vernetzt viele verschiedene Signalwege und kann hunderte Proteine wie z.B. TFs aktivieren

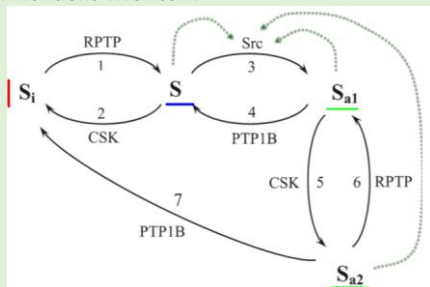
Src-Kinase besitzt **zwei verschiedene**

Phosphorylierungspositionen (inhibierend und aktivierend)

→ kann in vier verschiedenen Aktivitätsformen vorliegen

→ Aktivitätsformen sind nicht unabhängig voneinander

→ **NICHT Michaelis-Menten!**

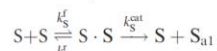


Inaktiv, partiell aktiv, aktiv, Autophosphorylierung

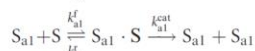
→ hat 4 Zustände (Mass-action-Kinetik), 7 Reaktionen und kann sich **Autophosphorylieren** (positives Feedback)!

□ S_i : inhibited (pY527, Y416)

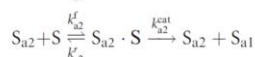
□ Autocatalytic network:



□ S : partial act. (Y527, Y416)



□ S_{a1} : active (Y527, pY416)



□ S_{a2} : active (pY527, pY416)

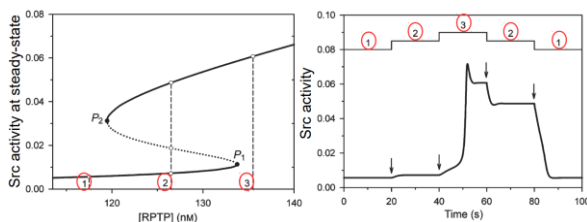
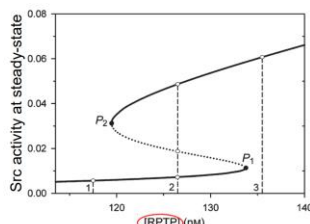
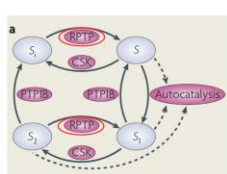
DYNAMISCHE VERHALTEN DER SRC-KINASE:

1. DYNAMISCHES GEDÄCHTNIS:

Alle drei aktiven Formen (S , S_{a1} und S_{a2}) sind zu

Autoaktivierung fähig → **positives Feedbacksystem** →

ermöglicht ein **dynamisches Gedächtnis**



RPTP: Phosphatase, welche die inaktive Form S_i durch Dephosphorylierung in die teilweise aktive Form S überführt
Tiefe [RPTP] → [S] tief (Aktivität der Src tief!)

Hohe [RPTP] → [S, S_{a1} , S_{a2}] hoch (Aktivität der Src hoch!)

mittlerer [RPTP] → SRC-Aktivität hängt von der vorherigen Aktivität ab!

Die Src-Kinase, ein Integrator-Protein, das an mehreren verschiedenen Signalwegen beteiligt ist, kann aufgrund von ihrem mehrstufigen Aktivierungszyklus verschiedene dynamische Verhaltensweisen zeigen, die erklären, wie sie unterschiedliche Zellantworten einleiten kann.

SENSITIVITÄTSANALYSEN

können Schlüsselkomponenten in Signalwegen identifizieren

Sensitivitätsanalyse: Man untersucht, wie sensitiv die Höhe des Outputs (z.B. der Aktivität eines Signalmoleküls) darauf reagiert, wenn die Parameter für die Menge verschiedener Moleküle variiert werden

→ Bestimmt die Menge eines Moleküls besonders stark die Höhe des Outputs, ist dieser sehr sensitiv für die Modulation der Molekülkonzentration → Durch Inhibition oder Aktivierung dieses Moleküls kann der Output sehr stark beeinflusst werden

ERBB-SIGNALWEG:

Interaktion: MAP-Kinase- und dem AKT-Signalweg, wenn sie von epidermal growth factors (EGF) aktiviert werden:

Signalwege werden durch eine komplexe Wechselwirkung von vier Rezeptoren (ErbB1 bis ErbB4; Hinweis: ErbB1 = EGFR) und mehreren Liganden der EGF-Familie aktiviert

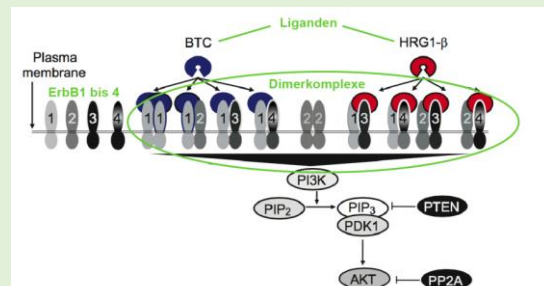
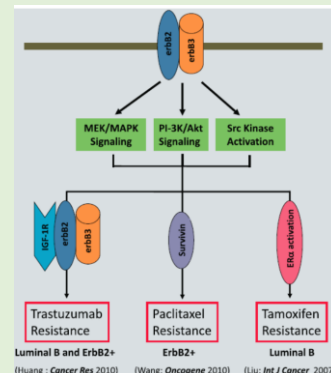


Abbildung 7.5: Schematische Darstellung des ErbB-Signalnetzwerks. Es gibt vier ErbB-Rezeptoren (1-4) an der Zellmembran, die den AKT-Signalweg durch Bildung von Dimeren aktivieren können. Zwei wichtige Liganden sind BTC und HRG1-β, beides Mitglieder der EGF-Familie.

ErbB2: keine Ligandenbindestelle

ErbB3: keine Kinaseaktivität → müssen Heterodimere bilden



Inhibition von ErbB3 hat die stärksten Erfolge gebracht

Output: Menge an phosphoryliertem AKT

Grosse dynamische Modelle von Signalnetzwerken können genutzt werden, um mögliche Ansatzstellen für Medikamente zu identifizieren. Besonders sensitive Signalmoleküle kann man durch eine Sensitivitätsanalyse ermitteln

ÜBUNGEN

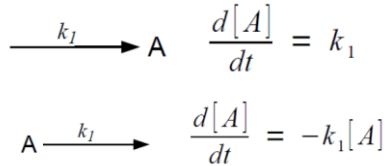
ÜBUNG 2

1) Was sind die **beiden Hauptannahmen** bei der kinetischen Modellierung von Stoffwechselreaktionen mit **gewöhnlichen Differentialgleichungen (ODEs)**?

1. Die Anzahl der Moleküle ist gross
2. Das System ist gut gemischt (keine räumlichen Heterogenitäten)

2) Schreiben von den jeweiligen Elementarreaktionen die ODEs auf

→ hin +), weg -) immer mal den Konzentrationen



3) Welche **zwei Annahmen** und **zwei Vereinfachungen** wurden getroffen, um eine analytische Lösung für die **Michaelis-Menten-Annäherung** abzuleiten?

- Keine Rückreaktion des Produktes P: [P] hängt nur von [ES] ab (Irreversible Reaktion)
- Enzym wird konserviert → Gesamtmenge an Enzym bleibt konstant: $[E]_{\text{total}} = [E] + [ES]$
- Das System befindet sich im quasi stationären Zustand: Die Menge an ES ist konstant: $d[ES]/dt \approx 0$
- Substratüberschuss S

4.) Berechnen sie K_m

$$K_m = \frac{k_{-1} + k_2}{k_1} = 20$$

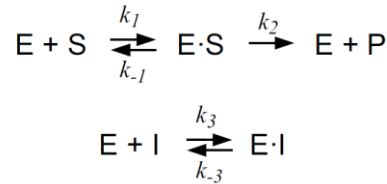
5) Bei Erhöhung der Start-Enzymkonzentration

- erhöht sich $V_{\max}$
- K_m bleibt konstant

6.) Bei Erhöhung der kompetitiven Inhibitor-Konzentration –

- ändert sich $v_{\max}$ nicht
- K_m verändert sich (erhöht sich)
- Weniger [E] ist vorhanden

7) Leite die 'reaction rate expression (v)' zur kompetitiven Hemmung, unter den folgenden Annahmen ab



- Quasi Steady State assumption for intermediate complexes ES and EI: $\frac{d[ES]}{dt} = 0, \frac{d[EI]}{dt} = 0$
- Conservation of Enzyme (E) species: $E_T = E + ES + EI$
- $\frac{d[P]}{dt} = \text{reaction rate} = v$

Vorgehen:

1. Alle ODEs aufstellen
2. Merke: $[E]^T = [E] + [ES] + [EI]$
3. $D[ES]/dt = 0$ setzen und vereinfachen → nach [E]
Auflösen →

$$[E] = \frac{(k_{-1} + k_2)[ES]}{k_1[S]} \xrightleftharpoons[\text{We know: } k_m = \frac{(k_{-1} + k_2)}{k_1}]{k_m} [E] = k_m \frac{[ES]}{[S]}$$

4. $D[EI]/dt = 0$ setzen und vereinfachen

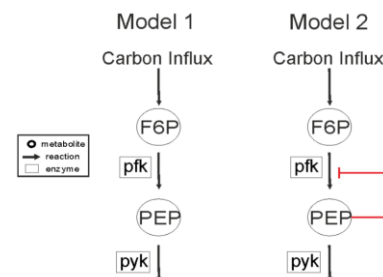
$$[EI] = \frac{[E][I]}{\frac{k_{-3}}{k_3}} \xrightleftharpoons[k_I = \frac{k_{-3}}{k_3}]{k_I} [EI] = \frac{[E][I]}{K_I} \xrightarrow{\text{Eq. 1}} [EI] = \frac{k_m[ES][I]}{K_I[S]}$$

5. Alles in '2' einsetzen und nach [ES] auflösen
6. [ES] in P einsetzen

$$v = \frac{d[P]}{dt} = k_2[ES] \xrightarrow{\text{Eq. 3}} \frac{d[P]}{dt} = \frac{k_2[E]^T[S]}{[S] + k_m \left(1 + \frac{[I]}{K_I}\right)} \xrightleftharpoons[v_{\max} = k_2[E]^T]{v_{\max} = k_2[E]^T}}$$

$$\Rightarrow \frac{d[P]}{dt} = v = \frac{v_{\max} * [S]}{[S] + K_m \left(1 + \frac{[I]}{K_I}\right)}, [S] \cong [S_0]$$

MICHAELIS-MENTEN KINETIK



8) Based on the reaction scheme in Figure 1, write the **Balance Equations** for the species' concentrations of Model 1 under the **assumption of Michaelis Menten kinetics** and a **constant Carbon Influx**.

→ 1. Darstellung durch v

$$\rightarrow 2. \text{ Michaelis-Menten: } v = \frac{v_{\max} * [S_0]}{[S_0] + K_M}$$

$$\frac{d[F6P]}{dt} = v_{\text{carboninflux}} - v_{\text{pfk}} = v_{\text{carboninflux}} - \frac{v_{\max-\text{pfk}}[F6P]}{K_{M-\text{pfk}} + [F6P]}$$

$$\frac{d[PEP]}{dt} = v_{\text{pfk}} - v_{\text{pyk}} = \frac{v_{\max-\text{pfk}}[F6P]}{K_{M-\text{pfk}} + [F6P]} - \frac{v_{\max-\text{pyk}}[PEP]}{K_{M-\text{pyk}} + [PEP]}$$

9) write the reaction rates of the irreversible and reversible reactions. **Irreversible reactions follow the Michaelis Menten law and reversible reactions follow the mass action law**

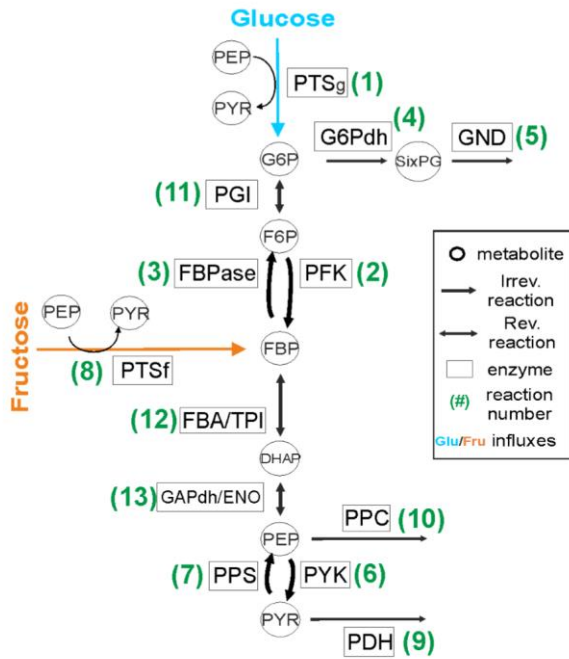


Figure 1: Reaction scheme of Glycolysis

(Glucose Influx) PTSg*: V_{max1}

(Fructose Influx) PTSf*: V_{max8}

* = Influxes (values known)

PFK (Irreversibel): $V_{max2} * \frac{[F6P]}{[F6P] + K_{m2}}$

FBPase (Irreversibel): $V_{max3} * \frac{[FBP]}{[FBP] + K_{m3}}$

PGI (Reversibel):

$V_{max11forward} * [G6P] - V_{max11backward} * [F6P]$

GAPDH/ENO (Reversibel):

$V_{max13forward} * [DHAP] - V_{max13backward} * [PEP]$

REMEMBER:

- The initial concentrations of metabolites do **not have an effect on the steady state values** of the metabolites, but only on their **dynamics UNTIL they reach a steady state**.
- the Glucose Influx value has an effect on the steady state values of the metabolites → the higher the influx, the higher the steady-state concentrations

9) What is the **main assumption** of FBA?

The system is in **steady state**: for each metabolite, sum of input fluxes = sum of output fluxes

10) What are the **3 inputs** required to perform FBA of a metabolic network?

- stoichiometric matrix
- reaction constraints
- optimization function

11) Can we introduce only one modification to the model, ATP consumption by the reaction R1? If not, what is necessary in order to simulate the model and why?

The model always **need to be balanced**. When adding **ATP as a compound** you need to make sure to have **production of ATP to counterbalance its consumption**.

12) Write a pseudocode that search for all possible of synthetic lethal pairs of enzymes

Model → Stoichiometric model of E. coli metabolism

For i=1:N (For each enzyme in the model)

Indx_rxn_A = find reactions catalyzed by the enzyme i;

For j=1:N (For each enzyme in the model j~i)

Indx_rxn_B = find reactions catalyzed by the enzyme j;

Model_KN = Model (temporary model variable to be modified by double gene deletions)

Model_KN.lb(Indx_rxn_A, Indx_rxn_B) = 0
(force lower bounds of the reactions catalyzed by enzyme i and j to be 0)

Model_KN.ub(Indx_rxn_A, Indx_rxn_B) = 0
(force upper bounds of the reactions catalyzed by enzyme i and j to be 0)

Optimize biomass production for Model_KN

Save the solution of maximum biomass production

end

end

Find all pairs of enzyme for which maximum biomass production is equal to 0 (synthetic lethal enzyme pairs)

PRÜFUNGSAUFGABE

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. (Stichworte reichen aus)

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

Differentialgleichungen → Nicht FBA, da wir nicht im GGW befinden → Störungseinfluss nicht steady state

b) Welche Informationen und welche Daten brauchen Sie dafür? (1.5 P)

- dynamische Konzentrationen: Metabolitkonzentration (Enzymkonzentration)
- Enzymatische Parameter
- Wo interagiert der Inhibitor und wie sieht der Stoffwechselweg insgesamt aus

SYSTEMBIOLOGIE TEIL 2-

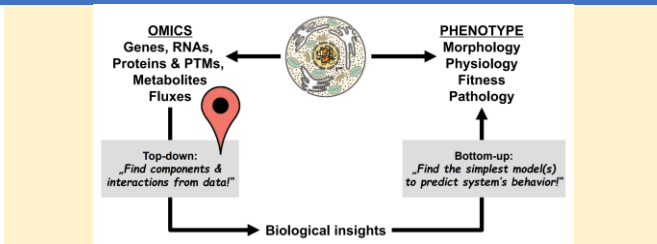
DATENANALYSE

GROSSE DATENMENGEN

5 Hauptmethoden, um mit grossen Datenmengen umzugehen:

- Feature selection
- Univariate analysis of large datasets
- Multivariate analysis
- Clustering
- Link inference

OMICS-DATEN

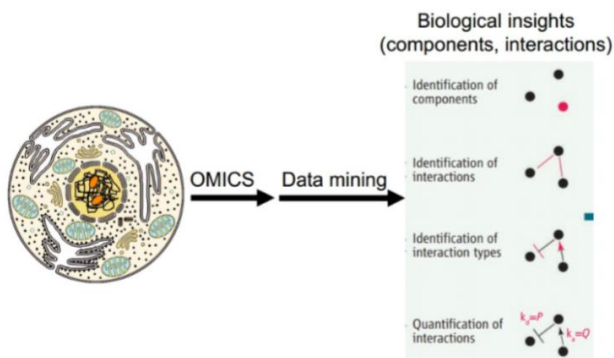


Bottom-up → Man will ein System charakterisieren und herausfinden, wie es funktioniert → Aus bekannten Erkenntnissen wird versucht ein Modell zu finden, welches das Verhalten des Systems am besten/einfachsten voraussagt. Aus den Ergebnissen lassen sich Voraussagen über den Phänotyp machen

Top-down → Aus Zellen werden Omics-Daten extrahiert, die verwendet werden, um Komponenten und Interaktionen zu finden, die biologische Einblick ermöglichen

What do we expect to obtain from omics data?

Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms



What are the typical problems associated with omics data?

PROBLEME BEI DER ANALYSE VON OMICS-DATENSÄTZE

- Viele Features aber nicht so viele Proben/Samples
- Daten sind oft verrauscht
- Experimente sind nicht reproduzierbar

→ Nicht reproduzierbare Experimente führen oft zu schlechten Vorhersagen

- Risiko: Overfitting der Daten

→ Gute Statistik nötig!

Overfitting: Einbeziehung des Rauschens in die Daten. Ein 'overfitted Model' benutzt demnach kaum reproduzierbaren Daten → Solche Modelle führen zu schlechten Vorhersagen über neue Datensätze

SIGNIFIKANZ

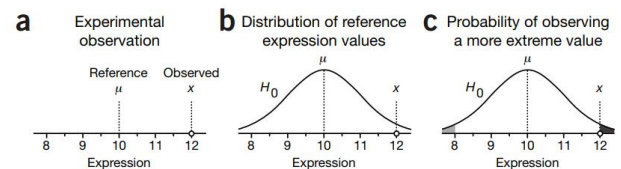


Figure 1 | The mechanism of statistical testing. (a–c) The significance of the difference between observed (x) and reference (μ) values (a) is calculated by assuming that observations are sampled from a distribution H_0 with mean μ (b). The statistical significance of the observation x is the probability of sampling a value from the distribution that is at least as far from the reference, given by the shaded areas under the distribution curve (c). This is the P value.

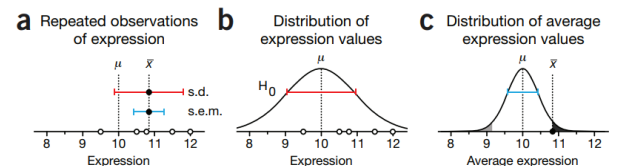


Figure 2 | Repeated independent observations are used to estimate the s.d. of the null distribution and derive a more robust P value. (a) A sample of $n = 5$ observations is taken and characterized by the mean $\bar{x}$, with error bars showing s.d. (s_x) and s.e.m. ($s_x/\sqrt{n}$). (b) The null distribution is assumed to be normal, and its s.d. is estimated by s_x . As in **Figure 1b**, the population mean is assumed to be μ . (c) The average expression is located on the sampling distribution of sample means, whose spread is estimated by the s.e.m. and whose mean is also μ . The P value of $\bar{x}$ is the shaded area under this curve.

Statistics quantifies the **outlier status** of an observation by the probability of sampling another observation from the null distribution that is as far or farther away from μ . This probability is the P value, which is the output of common statistical tests.

'prosecutor's fallacy': P value is often misinterpreted as the probability that the null hypothesis (H_0) is true we assumed that H_0 is true and a small P value (for example, $P = 0.05$) only tells us that an improbable event has occurred in the context of this assumption. The degree of improbability is evidence against H_0 and supports the alternative hypothesis that the sample actually comes from a population whose mean is different than μ . Statistical significance suggests but does not imply biological significance

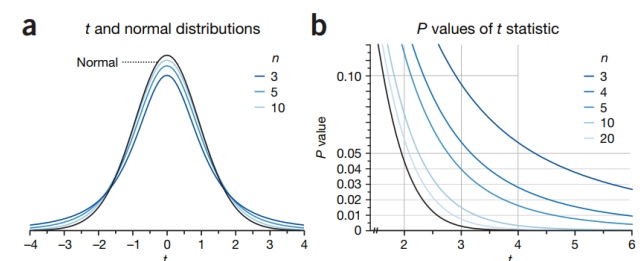


Figure 3 | The t and normal distributions. (a) The t distribution has higher tails that take into account that most samples will underestimate the variability in a population. The distribution is used to evaluate the significance of a t statistic derived from a sample of size n and is characterized by the degrees of freedom, d.f. = $n - 1$. (b) When n is small, P values derived from the t distribution vary greatly as n changes.

STATISTISCHE TESTS

Statistical significance **suggests but does not imply biological significance**

What is the role of statistics?

- Statistik ist wichtig, aber nicht ausreichend
- Studien beginnen mit dem **P-Wert**
- ACHTUNG: nicht beim p-Wert stehenbleiben → an den **Kontext** und die **Annahmen** denken!
- Nicht mit P-Wert enden → unzureichend!
- Man sollte immer noch **experimentell überprüfen (validieren!)**
- Zudem: **Positive und negative Kontrollen** sind sehr wichtig!
- Hypothesen als Anfang für die Forschung, kein Endpunkt
- Die Forschung wird nicht voranschreiten, wenn sie nur auf absolut bewiesenen Tatsachen beruht. Es ist in Ordnung, unsichere Hypothesen zuzulassen, solange man sich dessen bewusst ist.
- Biologen nutzen Data-Mining zur Erstellung von Hypothesen. Einige davon sind sicher falsch → Diese Hypothesen sollen darum den **Start** für die Forschung bedeuten, **nicht den Endpunkt**. Hypothesen müssen also **mittels follow-up Experimenten getestet werden**.
- Gute Statistiken sind ein Indikator für gute Laborarbeit (Reproduzierbarkeit).

BEST PRACTICE IN (BIO) DATA MINING

1. Start from a **biological question/problem!**
- Don't expect an answer at all costs! → lack of data, noise, lack of response, wrong question
2. Always think about **positive and negative controls**
3. It's a good habit to (try to) **keep data mining simple!**
4. Test different approaches (Ansätze) and parameters
5. Focus on **reproducible results**

ZWEI GRUPPEN PROBLEM

Aufgabe: 2 Samples miteinander (versch. Klassen) vergleichen und signifikante Unterschiede finden:

UNIVARIATE ANALYSIS

- man betrachtet jeweils 1 Merkmal/Feature und vergleicht es der beiden Gruppen → Analyse wird für alle Merkmale unabhängig voneinander wiederholt → verwendet low-dimensional Data

VORGEHENSWEISE

1. Calculate magnitude of change (Größe der Veränderung)
 - Ist der Unterschied **relevant?**
- **fold-change** zwischen Gruppen Mittelwerten o. Mediane

$$FC = \frac{\text{mean(Gruppe A)}}{\text{mean(Gruppe B)}}$$

Für bessere (symmetrische) Visualisierung:

$$\log_2(FC) = \log_2 \left(\frac{\text{mean(Gruppe A)}}{\text{mean(Gruppe B)}} \right)$$

2. Verify statistical significance

- Ist dieser Unterschied statistisch signifikant?

→ **P-Wert** für die Nullhypothese

Nullhypothese: Beide Gruppen sind gleich

GES: Wahrscheinlichkeit, dass die gesammelten Daten von gleichen Gruppen abstammen → Nullhypothese wahr

Überprüfung der statistischen Signifikanz durch:

- T-Test
- Mann-Whitney/Wilcoxon/Ranksum-Test
- Permutationstest

→ Alle Tests liefern einen P-Wert

P-WERT:

Der p-Wert entspricht **der Wahrscheinlichkeit, die Resultate zu erhalten, wenn die Nullhypothese stimmt**

ACHTUNG: Er entspricht nicht der Wahrscheinlichkeit, dass die Nullhypothese (keine Unterschiede) stimmt

('prosecutor's fallacy')

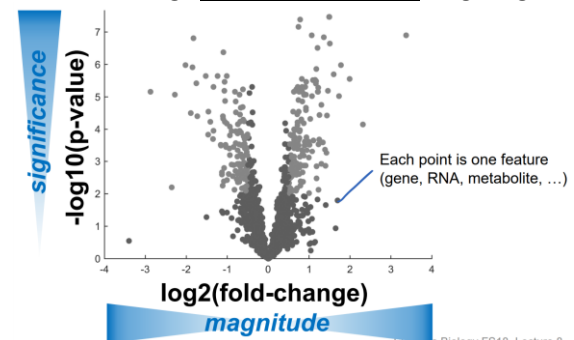
- In der Biologie gilt ein p-Wert von 0.05 als akzeptabel, nice to have ist 0.01

3. Daten werden als Volcanoplot visualisiert:

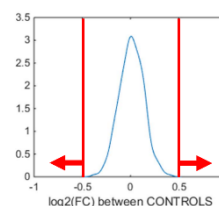
Methode, um den Fold-Change und die statistische

Signifikanz eines Zwei-Gruppen-Vergleichs zu evaluieren

→ Für jedes Feature (z.B. ein Gen) wird der p-Value gegen den Fold-Change doppellogarithmisch aufgetragen

THRESHOLDS (SCHWELLENWERT)Foldchange Treshold:

- Wahl eines zufälligen Wertes (aus Erfahrung)
- **Empirische Vorgehensweise:** Man nimmt eine **Kontrollgruppe, die keine Unterschiede aufweisen sollte** → Man testet alle möglichen Kombinationen → Was für Unterschiede kann man finden, obgleich sie diese in derselben Gruppe befinden → **threshold** Bestimmung → Wert wird also so gewählt, dass Dinge, die identisch sind, nicht als signifikant unterschiedlich bezeichnet werden

P-Wert Threshold:

Traditionally in biology:

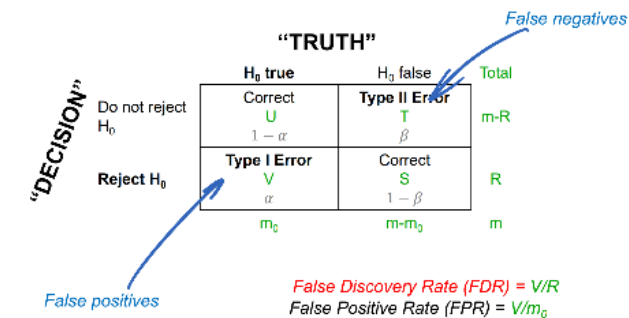
p-value < 0.05 : acceptable!

p-value < 0.01 : nice to have!

PROBLEM DES MULTIPLLEN TESTEN

→ Bei einer grossen Anzahl von Tests wird die Wahrscheinlichkeit von Falsch-Positiven Ergebnissen sehr gross
 Man bestimmt ein alpha ($\alpha=0.05 \rightarrow 5\%$ false positive) →
 PROBLEM: α erhöht sich mit zunehmender Stichprobe →
 hat sehr viele falsche Aussagen → $FPR = \alpha * N$
 Probability of making at least 1 error → $1 - (1 - \alpha)^m$
 → P-Wert muss angepasst/korrigiert werden!

Wenn viele Einzeltests zu einem gegebenen α -Niveau (also der Fehler erster Art; in der Regel 5%) durchgeführt werden, erhöht sich die Chance für die Gesamtzahl der Hypothesen zufällig ein signifikantes Ergebnis zu bekommen →
 "Alphafehler-Kumulierung" → $\alpha * N \rightarrow$ Fehler summieren sich bei mehreren durchgeführten Hypothesen stark auf, so dass das Signifikanzniveau von 5% stark überschritten wird



- $\alpha = 0.05$ (~Type I error, FPR) → we expect a 5% chance of making a false positive inference
- The false positive rate (FPR) is the chance of inferring an effect when no effect is present
- The false discovery rate (FDR) is the fraction of positive inferences that are false
- power is the fraction of real effects that are detected

LÖSUNGEN GEGEN MULTIPLES TESTEN:

FAMILY WISE ERROR RATE (FWER):

a) Bonferroni-Methode:

P-Wert wird mit der Anzahl Tests m multipliziert (oft zu strikt!) → ineffizient → reduziert die Macht zu stark! → In der Biologie kann man einige Falsch-Positive tolerieren)

$$\text{reject } H_i \text{ if } p_i \leq \frac{\alpha}{m}$$

FALSE DISCOVERY RATE (FDR): (weniger strikt als FWER)

b) Control False Discovery Rate (Benjamin-Hochberg)

Control False Discovery Rate [Benjamini and Hochberg 1995]

Controls the proportion of false positives among the set of rejected hypotheses

p-values > FDR-adjusted p-values, reject H_i if $\text{adj. } p_i \leq \alpha$

OK if differences are rare ($p_0 \approx 1$)

c) Control Positive False Discovery Rate

Verwenden, wenn es sehr viele Unterschiede gibt (hier braucht man aber sehr viele features!) → Q-Werte

Control positive False Discovery Rate [Storey and Tibshirani 2003]

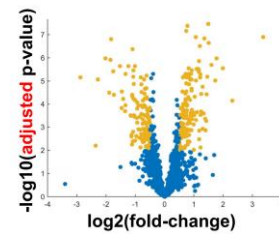
Controls the rate that discoveries are false (pFDR). Uses all p-values to derive q-values (minimum FDR that can be attained when calling that feature significant)
 example: if gene X has $q = 0.02$ it means that 2% of genes that show p-values at least as small as gene X are false positives.

Better if differences are frequent ($p_0 < 1$)

A test's q value is the minimum FDR at which the test would be declared significant. This FDR value is a collective measure calculated across all tests with $FDR \leq q$

Wenn ich bei der FDR z.B. 100 signifikante Ergebnisse habe kann man bei einem Signifikanzniveau von $\alpha=0.05$ davon ausgehen, dass davon im Mittel $100 * 5\% = 5$ Hypothesen fälschlicherweise abgelehnt wurden, während bei der FWER 95% Sicherheit besteht, dass alle signifikanten Ergebnisse korrekt sind.

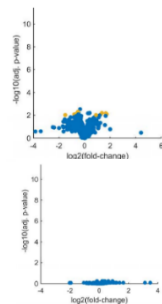
→ MERKE: Man verwendet immer einen **adjusted p-value!**



Gelb = signifikant

VERDÄCHTIGE MESSUNGEN:

Ziel: Man möchte robuste Daten



Verändert man hier den Threshold (p-Wert) nur minimal würde es eine komplette Veränderung geben

→ Es würden sehr schnell sehr viele Punkte signifikant werden (gelb) → keine robuste Daten!

rechte Punkte → Hoher fold-change aber keinen signifikanten P-wert

Was passiert, wenn die univariate Analysis ... hervorbringt	
keine signifikanten Änderungen	Datenqualität überprüfen Anzahl Replikate erhöhen
nur wenige signifikante Änderungen	Das ist gut!
sehr viele signifikante, aber kein klares Ranking	Ist üblich, aber schwierig daraus Hypothesen zu generieren Kombination von primären und/oder sekundären Effekten → Enrichment Analyse

ENRICHMENT ANALYSIS (EA)

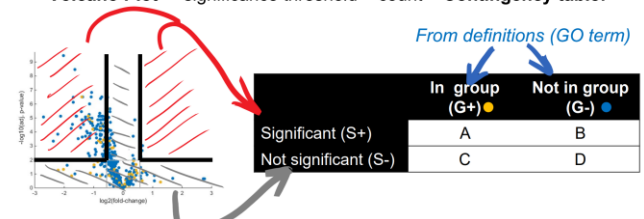
Die Enrichment-Analyse wird angewendet, wenn bei der univariaten Analyse **viele signifikante Features** zwischen zwei Gruppen gefunden wurden (Falls mehrere Features Signifikant sind, kann keine Hypothese formuliert werden)

Ziel: Herausfinden, ob unter allen signifikant veränderten Features (z.B. Gene, Proteine, Metaboliten) eine bestimmte Gruppe übermässig stark vertreten ist

Volcano Plot → significance threshold → count →

Contingency table (Fisher's Exact Test durchführen)

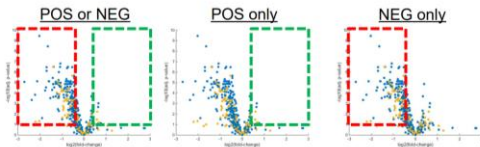
Volcano Plot > significance threshold > count > Contingency table:



GENE SET ENRICHMENT ANALYSIS (GSEA)

GSEA ist eine Methode, um Klassen von Genen oder Proteinen zu identifizieren, die in einem grossen Set von Genen überrepräsentiert sind und einen Einfluss auf einen bestimmten Phänotyp haben

1. Decide on what signs do we want to include:



2. Procedure for 1 pathway/group/process:

1. Choose very permissive thresholds (e.g. $|\log_2FC| > 0.2$ and $p\text{-value} < 0.1$)
2. Rank all hits on either p-value (preferred) or $\log_2FC$
3. Build contingency tables with either 2, 3, 4, ..., ALL top hits
4. For each table, calculate p-value using Fisher's exact test
5. Keep lowest p-value = "best" enrichment

3. Repeat for all pathways

4. FDR-correction (Benjamini-Hochberg or Storey)

Gen Set mit dem höchsten $-\log_{10}(P\text{-Wert})$ hat den höchsten Enrichment → Peak im GSEA output plot

SELF-CHECK

Was, wenn kein Enrichment gefunden wurde (obgleich hunderte von Genen unterschiedlich waren?)

Kommt häufig vor → gibt eventuell ganz viele Anreicherungen, sodass keine wirklich mehr signifikant ist → Das ist kein Hinweis auf schlechte Daten. Eventuell gibt es aber ungetestete Gruppen.

Was passiert, wenn man eine Enrichment-Analyse ohne Definierung des $\log_2FC$ oder des p-value Thresholds durchführt?

Threshold wird definiert als: $|\log_2FC| > \dots$ und $p\text{-value} < \dots$
Verändert sich nicht → Wenn man keine Grenzwerte für den FC oder p-Wert definiert, dauert die Analyse einfach länger (könnte man weglassen)

SUMMARY

- Bei Data-Mining und Omics-Daten geht es um die **Generierung von Hypothesen**
- Wenn man auf absoluter Sicherheit beharrt, wird man wahrscheinlich testbare Hypothesen nicht entdecken können
- Im einfachsten Ansatz werden signifikante Features durch **univariate Analyse aller Features** (unabhängig) gefunden → **t-Test** → FDR-Korrektur
- **Signifikant veränderte Prozesse werden durch EA gefunden** → Rangierung nach p-Wert → Subsets testen → Fisher's Exact Test → FDR-Korrektur

FEATURE SELEKTION

Es geht darum, in einem biologischen System diejenigen Komponenten zu finden, die einen bestimmten Output/Phänotyp/Funktion beeinflussen



Durch Feature Selektion möchte man herausfinden, welche Variablen (am meisten) mit einem Phänotyp (oft Krankheiten) korrelieren

- Welche Veränderungen im menschlichen Genom korrelieren am meisten mit dem Auftreten erblichen Krankheiten?
- Gibt es einzelne Veränderungen im Genom, die stark mit dem Auftreten solcher erblichen Krankheiten korrelieren?

WARUM FEATURE SELEKTION NOTWENDIG IST:

- Um ein besseres Verständnis der molekularen Mechanismen zu erhalten, die mit einem bestimmten Phänotyp korrelieren oder mit diesem sogar kausal zusammenhängen → Besseres Kausalverständnis gewinnen → welche Gene beeinflussen Phänotyp, der uns interessiert?
- Mit Feature Selection kann man die Dimensionen reduzieren um Modelle zu generieren, die **einfacher** und **günstiger** zu behandeln sind → **niedrigdimensionale Modelle** erhalten
- Anzahl Gene, die eine Krankheit beeinflussen hinunterbringen → Vorbereitungsschritt
- **Verrauschte/ Irrelevante/ Fehlerhaft gemessene** Features werden so aus dem Datensatz **entfernt (reduziert Overfitting)**

UNIVARIATE FEATURE SELEKTION

Bei der univariaten Feature Selection werden die Features jeweils isoliert betrachtet → **Jedes Feature erhält einen Score** (via pearson correlation oder Mutual Information) und wird nach diesem sortiert (Höher ist besser) (Score = Stärke der Korrelation zwischen Feature und Antwort)

Ermöglicht ein besseres Verständnis der Daten

Optimiert die Features nicht für Generalisierungen

VORGEHENSWEISE:

1. Für jedes Feature j wird der Score $r(j)$ berechnet
2. Die Features werden gemäss ihrem Score geordnet und als Liste zurückgegeben

$$\arg \max_j (r(\mathbf{x}(:, j), \mathbf{y}))$$

$\mathbf{y}$: Phänotypen aller Teilnehmer

J : Features

$\mathbf{X}(:, j)$: Vektor, der die **Expressionslevels von Gen j aller Patienten** beinhaltet (Datenmatrix → Individuen*Features)

r : Korrelationskoeffizient; Stärke von der statistischen Abhängigkeit/ Korrelation

→ Welche Spalte dieser Matrix korreliert am meisten mit dem Phänotyp ($\mathbf{y}$)? → Wir wollen j (Features) finden, das am meisten korreliert → an der die Funktion r also ihr **Maximum** besitzt (Gen mit dem höchsten Score)

Zwei Informationskriterien finden Anwendung:

Pearsons Correlation Coefficient:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

r_{xy} als Mass für die lineare Beziehung → gibt an, ob sich Features und Phänotyp gemeinsam bezüglich des Mittelwertes verändern

nahe 1 → hohe Korrelation

negativ → Antikorrelation (hohe Werte eines Vektors mit tiefen Werten des anderen Vektors korreliert)

0 → keine Korrelation → Veränderungen sind zufällig

Mutual Information:

$$I(\mathbf{x}, \mathbf{y}) = \sum_{z_x \in Z_x} \sum_{z_y \in Z_y} p(z_x, z_y) \log \left(\frac{p(z_x, z_y)}{p(z_x)p(z_y)} \right)$$

- Wie häufig treten sie gemeinsam auf?
 - Häufigkeit eines gemeinsamen Auftretens mit dem einem zufälligen, gemeinsamen Auftreten vergleichen
 - $p(z_x), p(z_y)$: Wahrscheinlichkeit des zufälligen Auftretens der einzelnen Werte
 - $p(z_x, z_y)$: Wahrscheinlichkeit des zufälligen gemeinsamen Auftretens
- Bei hoher Korrelation wird I gross → I wird grösser, falls Wahrscheinlichkeit eines gemeinsamen Auftretens häufiger ist als ein zufällig gemeinsames Auftreten

PITFALLS

1. Grösse/Umfang der Lösung (Lösungsvektors)
2. Multiples Hypothesentesten
3. Instabilität

1. SIZE OF SOLUTION

Wie viele Features (Gene, SNPs) wollen wir finden?

LÖSUNG:

1.1 Probe Method: Zufalls Features generieren

→ Man generiert zufällig ein Störfeature z und wählt alle Features j , für die gilt: $r(j) > r(z)$

1.2: Significance Method: Statistik, P-Wert

Man berechnet ein p-Wert für die Beziehung zwischen jedem Feature und dem Phänotyp. Dann werden alle Features ausgewählt, deren Assoziation signifikant ist.

2. MULTIPLE HYPOTHESES TESTING

p-Wert als Wk., eine solch starke Assoziation zu finden, falls kein Zusammenhang zwischen den Variablen besteht

→ Wk. für einen signifikanten Befund, falls keine wahre Assoziation zwischen Variablen (Genotyp) und Phänotyp besteht

- oft signifikanter falls p-Wert $< \alpha$ (α oft 0.05 oder 0.01)

Wiederholtes Testen führt dazu, dass man falsch positive Korrelationen findet → **Problem des multiplen Testen**

LÖSUNG:

Hypothesentesten! → Korrektur mit z.B. Bonferroni (α/n)

(ACHTUNG, oft zu strikt, findet fast keine signifikanten Gene mehr!) → geeigneter: False Discovery Rate Korrektur

INSTABILITY

Instabilität = kleine Änderung der Daten oder des Kriteriums führt zu grosse Änderung des Rankings

→ Wenn Feature Selection auf verschiedene Subsets oder mit unterschiedlichen Ranking-Kriterien durchgeführt wird, kann die Outputliste/ das Ranking der Features **stark variieren** und es wird **kein definitives, einmaliges Resultat erreicht**

- Datensätze mit **wenigen Patienten n** und **vielen Genen d** produzieren häufig instabile Resultate
- weitere Ursache für die Instabilität ist die **grosse Anzahl Methoden**, die es für Feature Selection gibt

LÖSUNG:

Ist eine Assoziation über mehrere Studien hinweg signifikant?

Stabilität analysieren → Feature-Rankings sollte mit verschiedenen Kriterien und mit verschiedenen Untermengen (leicht modifizierte Datasets) berechnet werden!

→ Experiment unter versch. Umständen wiederholen

- Zudem sollte ein Ranking nicht als einziges, definitives Resultat betrachtet werden!

→ Zwei populäre Strategien zum Zusammenführen der Resultate aus verschiedenen Rankings:

1. Den **Mittelwert/durchschnittliche Ranking** jedes Feature von allen Rankinglisten/Experimenten berechnen
2. Für jedes Feature die Wahrscheinlichkeit berechnen, dass es unter den Top-k ist (in jedem Experiment) (→ In wie vielen Wiederholungen lag ein Gen in den Top-K Bereichen des Experiments?)

LIMITATION DER UNIVARIANTEN F-S

Univariate Features: betrachtet Gene einzeln, unabhängig voneinander → **ist limitiert!** → kann keine Interaktiven, multiplikativen oder additiven Effekte betrachten!

+) ist sehr viel schneller, billiger und leichter zu interpretieren

+) weniger Parameter bestimmen (kein lambda)

- Berücksichtigt **nur den Effekt von einzelnen Genen**
- Berücksichtigt **keine Korrelation** zwischen Genen
- Berücksichtigt **keine additiven Effekte** zwischen Genen
- Berücksichtigt **keine Interaktionen** zwischen Genen

Im Prinzip wird also der Kern der Systembiologie ignoriert

→ Alternative zu **univariate** Features:

MULTIVARIATE FEATURE SELECTION -

«Verwenden Sie die wichtigsten Gene, um den Phänotyp anhand der Genexpressionsstufen vorherzusagen»

Multivariate Modelle betrachten **mehr** als nur eine isolierte Variable:

- können besser die Wirklichkeit widerspiegeln
- Gruppeneffekte können untersucht werden
- gegenseitige Beeinflussungen werden berücksichtigt
→ Gene können z.B. auch additive Effekte (verstärken sich gegenseitig) oder multiplikative Interaktionen (Phänotyp tritt nur auf, wenn mehrere Gene mutiert sind) besitzen
- Verschiedene Modelle: Additives Modell, Lasso Model, Ridge Regression, Elastic Net

Ziele der multivariaten Feature Detection:

1. relevante Features identifizieren
2. auf der Basis von Genexpressions-Levels Aussagen über den Phänotyp machen

Verwendet Modelle mit linearer Regression

Man macht eine Linearkombination jedes Features

$$y = \sum_{i=1}^d x_i \beta_i + \epsilon.$$

- Features bekommen ein **Gewicht von β** (ist gesucht)
- Einträge von β sagen etwas darüber hinaus, wie relevant ein Feature ist (β von 0 = irrelevant)
- y wird versucht mit x vorherzusagen

ADDITIVE MODEL

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

ZIEL:

- $X\beta$ soll möglichst nahe an y liegen (Fehler/euklidische Distanz soll minimiert werden)
- **Arg min → wir wollen die Gewichte (β) finden, die diesen Fehler minimieren**

Nachteil: oft keine Einträge (β) von 0 → man braucht einen **zusätzlichen Regulationsterm (λ)** (der Overfitting verhindert)

LASSO MODEL - L1-Norm

Lösungen mit wenigen $\beta \neq 0$ werden belohnt

$$\arg \min_{\beta} ||y - X\beta||_2^2 + \lambda_1 ||\beta||_1$$

- L1 Norm von β (=Betragssumme) wird minimalisiert

$$||\beta||_1 = \sum_{i=1}^d |\beta_i|$$

- Vorteile: Lösung β mit vielen 0 (verhindert Overfitting)
- Nachteil: Wenn es Gruppen von korrelierten Features gibt, wählt das Lasso-Modell nur 1 aus der Gruppe aus

RIDGE REGRESSION - L2-NORM

Lösungen, in denen korrelierte Features ähnliche Gewichtung erhalten, werden belohnt

$$\arg \min_{\beta} ||y - X\beta||_2^2 + \lambda_2 ||\beta||_2^2$$

- Dies wird durch **Minimierung des L2-Terms** erreicht

$$||\beta||_2 = \sqrt{\sum_{i=1}^d \beta_i^2}$$

- Vorteil: gleiche Gewichtung bei korrelierten Features
- Nachteil: Viele Einträge sind nicht 0

ELASTIC NET - L1/L2-NORM KOMBINATION

- Beide Normen werden minimiert:

$$\arg \min_{\beta} ||y - X\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$$

Beim Elastic Net-Modell werden die Vorteile aus dem Lasso Modell (**wenig non-zero**) und der Ridge Regression (**gleiche Gewichtung bei korrelierten Features**) kombiniert

Nachteil: Haben nun 2 Parameter, die wir in unserem Optimierungsproblem definieren müssen

SZENARIEN:

Szenario 1: grössere β Werte

Scenario 1 - Higher values: We compare two solutions $\beta = 0.5$ and $\beta^* = 0.4$.

- The L1-norm of β is 0.5, the L1-norm of β^* is 0.4 (Difference is 0.1).
- The squared L2-norm of β is 0.25, the squared L2-norm of β^* is 0.16.
- Changing the solution from β to β^* leads to an **improvement of 0.1 in terms of L1-norm, and of 0.09 in terms of squared L2-norm.**

L1-Norm1: Beta=0.4 ist um 0.1 besser als Beta*=0.5

L2-Norm: Differenz: 0.09

Szenario 2: Betas, die näher bei 0 liegen

Scenario 2 - Lower values: We compare two solutions $\beta = 0.2$ and $\beta^* = 0.1$.

- The L1-norm of β is 0.2, the L1-norm of β^* is 0.1 (Difference is again 0.1).
- The squared L2-norm of β is 0.04, the squared L2-norm of β^* is 0.01.
- Changing the solution from β to β^* leads to an **improvement of 0.1 in terms of L1-norm, and of 0.03 in terms of squared L2-norm.**

L1-Norm: Differenz: 0.1

L2-Norm: Differenz: 0.03

→ Bei der L1-Norm ist die Belohnung, die wir von der Änderung B auf B* erhalten, immer gleich! (bei gleicher Gewichtsreduktion) → belohnt das Absenken von hohen Werten gleichermassen wie das Absenken von kleinen Werten!

→ L2-Norm: Belohnt das Absenken **nicht gleichermassen!** Belohnt es, grosse Werte kleiner zu machen, aber weniger, kleine Werte noch kleiner zu machen → L2-Norm gewährleistet nicht, viele auf 0 abzusenken

PITFALL: OVERFITTING

Selection Bias führt zu Overfitting (liefert zu viele optimistische Resultate)

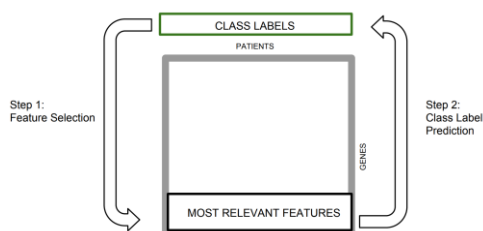
Overfitting kann umgangen werden, indem ein Trainings-dataset klar von einem Testdataset getrennt wird

Selection Bias: Geschieht, falls Feature Selection und Vorhersage auf dem gleichen Datensatz basieren

→ das Modell ist zu stark auf den betrachteten Datensatz angepasst und für andere Datensätze nicht zu gebrauchen

→ Ignorierung des Selection Bias führt zu übermässigen optimistischen Resultaten

UNSAUBERES VORGEHEN:



1. Schritt: Features auswählen, die stark mit dem Phänotyp korrelieren
 2. Schritt: genau diese Features werden gebraucht, um Vorhersagen über den Phänotypen zu machen → **Selection Bias**
- ich kann keine Aussage darüber machen, ob meine Features bei anderen Datensätzen zu guten Vorhersagen führen (Over-fitting)

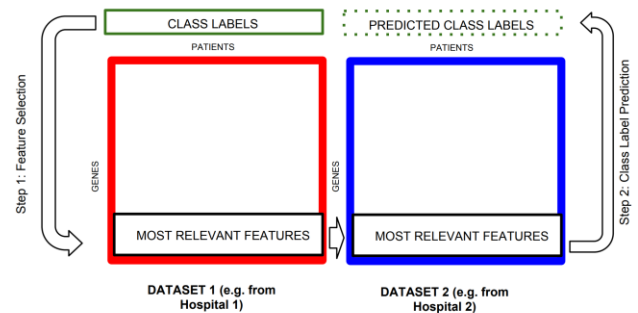
LÖSUNG

Feature Selection und Vorhersage dürfen nicht auf dem gleichen Datensatz basieren → Features werden aus Training-Sets ausgesucht und Vorhersagen werden auf einem separaten Test-Dataset gemacht

SAUBERES VORGEHEN:

Selection Bias vermeiden: Features werden nicht auf demselben Datensatz ausgesucht auf dem sie auch angewendet werden → Man nimmt 2 Datensätze oder unterteilt einen Datensatz in 2 Datensätze

1. **Datensatz** um mostrelevant Features zu bestimmen
2. **Datensatz**: um die Klassen vorauszusagen



→ Sauber zwischen Trainingsdaten and Testdaten (Klassifikator) unterscheiden

SUMMARY

PROBLEME BEI FEATURE SELECTION:

1. **Size of Solution**
 - Zufalls Features generieren oder P-Wert
2. **multiple hypothesis testing**
 - correct for multiple testing, at least using Bonferroni correction.
3. **instability** of most methods on sample datasets
 - Aggregate rankings from several subsamples of the data
 - Aggregate rankings for different methods
4. the **relative advantages and disadvantages of univariate and multivariate feature selection**, when choosing your method.
 - Univariate Methode: schneller, billiger und leichter zu interpretieren, weniger Parameter müssen bestimmt werden, berücksichtigt aber **nur den Effekt von einzelnen Genen und keine Korrelation, additiven Effekte oder Interaktionen** zwischen Genen
5. the problem of **overfitting**
 - Avoid it by cleanly separating the training dataset from the test dataset
 - Do not use the test set for feature selection

CLUSTERING

Als Clustering wird die **Suche nach Untergruppen ähnlicher Objekte** in einem gegebenen Datensatz bezeichnet.

Objekte in einer Untergruppe sollten einander ähnlicher sein als zu den anderen Objekten

- Auffinden von Genen, die z.B. über gemeinsame Transkriptionsfaktoren ko-reguliert werden
- Untergruppen von Patienten, welche an verschiedenen Varianten einer Krankheit leiden, bestimmen → Subphänotyp-Entdeckung

K-MEANS CLUSTERING

- **Centroid-basierte Clustering-Methode**
- beliebtester Clustering-Algorithmus
- Jeder Cluster wird durch einen Vektor repräsentiert (Dieser muss nicht notwendigerweise einem Punkt im Datensatz entsprechen)

ALGORITHMUS:

Input: Ein Set aus Datenpunkten $x_1, \dots, x_n$; Ein Integer k (Anzahl Cluster)

Output: Eine Partitionierung der Daten in k Cluster (ohne Überschneidung), sodass die folgende Funktion minimiert wird:

$$\sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|_2^2$$

μ_j : Mittelpunkt des j -ten Clusters.

ZIEL: Summe aller Abstände zwischen den Punkten und ihren jeweiligen Cluster-Mittelpunkten soll minimal sein

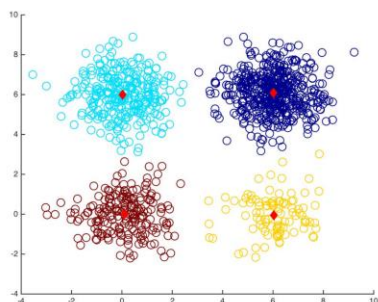
PROBLEM: schlimmstenfalls müssen alle Kombinationen durchgegangen werden → exponentiellen Aufwand

LÖSUNG: Verwendung des **Lloyds Algorithmus** → Computer such wiederholt nach **Gruppenmittelpunkten**

VORGEHEN:

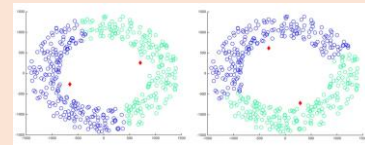
(bei k vermuteten Gruppen)

1. Am Anfang wählt er k zufällige Cluster-Mittelpunkte
2. Alle anderen Punkte werden nun demjenigen Mittelpunkt zugeordnet, dem sie am **nächsten** sind → führt zu k Gruppen
3. In jeder Gruppe wird der **neue Mittelpunkt gesucht** → k (neue) Mittelpunkte, mit denen dann wieder zu Punkt zwei zurückgekehrt wird
4. Wenn sich die Mittelpunkte nicht mehr ändern, wird die Berechnung beendet

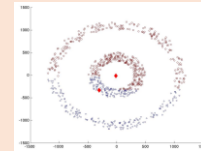


LIMITATIONEN

1. **Anzahl Cluster k** muss vom User selbst definiert werden
2. k-Means ist **Initialisierungsabhängig** → Je nachdem, wie die die Mittelpunkte am Anfang gewählt werden, resultieren andere Cluster



3. k-Means kann nicht-sphärische Cluster nicht erkennen



4. Die gefundene Lösung entspricht nur einem lokalen Optimum → möglicherweise existieren bessere Lösungen

SILHOUETTE PLOTS

k kann z.B. mittels Silhouette-Koeffizienten bestimmt werden. Dieser Koeffizient setzt folgende zwei Größen in Beziehung zueinander:

- $d(p, C)$: Durchschnittliche Distanz zwischen einem Punkt p und allen anderen Punkten im Cluster C
- $d(p, C')$: Durchschnittliche Distanz zwischen einem Punkt p und den anderen Punkten des zweit-nächsten Clusters C'

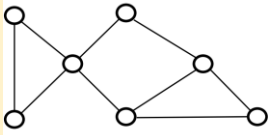
$$s(p) = \frac{d(p, C') - d(p, C)}{\max(d(p, C), d(p, C'))}$$

- $s(p)$ ist nahe bei 1, wenn ein Punkt klar in Cluster C lokalisiert werden muss.
- $s(p)$ ist nahe bei 0, wenn ein Punkt zwischen zwei Clustern liegt.
- $s(p)$ ist negativ, wenn der Punkt näher am anderen Cluster liegt.

GRAPH-BASIERTES CLUSTERING

ANNAHMEN:

1. Die Daten sind in Form eines Netzwerks/Graphen gegeben
2. Jeder Knoten ist ein Objekt
3. Kanten verbinden assoziierte Objekte
4. Die Gewichtung der Kanten widerspiegelt deren Länge, also die Distanz zwischen den Objekten

**ALGORITHMUS/ VORGEHEN**

1. Alle Linien mit einer Gewichtung, die grösser als der vom User definierte Grenzwert ist, werden entfernt
2. Alle verbundenen Komponenten im resultierenden Graphen werden identifiziert → Diese verbundenen Komponenten bilden ein Cluster



- Zwei Knoten gehören zur gleichen Grafikkomponente, wenn zwischen ihnen ein Pfad vorhanden ist

NACHTEIL/WEAKNESS→ Single-link effect

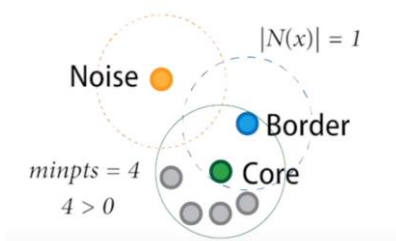
- Veränderte Werte der Gewichtung einer einzigen Kante (z.B. durch Messfehler, Rauschen) kann die Anzahl Cluster drastisch verändern (aus mehreren Clustern eines machen oder umgekehrt)
- Graph-based clustering ist also nicht robust!

DBSCAN-ALGORITHMUS*Density Based Spatial Clustering of Applications with Noise*

- **Noise robuste Variante** des Graph basierten Clustering

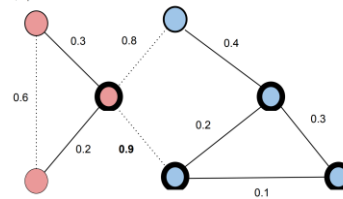
In DBSCAN gibt es drei Klassen von Punkten/Objekten:

- **Core Objects:** Ein Punkt wird als Core Object bezeichnet, wenn er eine bestimmte Anzahl Punkte (**Variable MinPts**) innerhalb einer **Distanz ϵ** hat
→ **MinPts** und ϵ sind benutzerdefinierte Parameter
- **Border Points:** Ein Punkt, der kein Core Object ist, aber innerhalb der Distanz ϵ von einem Core Object liegt
- **Noise:** Alle Punkte, die weder Core Objects noch Border Points sind

**ALGORITHMUS**

1. Wähle einen Punkt, der noch keinem Cluster zugeordnet ist
2. Wenn es Core Object ist (wenn sich MinPts in einer bestimmten Distanz von diesem Punkt befinden), füge ihn zu einem neuen Cluster C hinzu, wenn nicht, markiere ihn als Noise
3. Füge seine Nachbarn dem selben Cluster C hinzu
4. Überprüfe für jeden Nachbar, ob es ein Core Object ist
 - Wenn ja: Füge seine Nachbarn C hinzu und führe Schritt 4 auch für diese Objekte durch
 - Wenn nicht: Vergrößere das Cluster nicht weiter durch diesen Punkt
5. Starte wieder bei Schritt 1 bis alle Punkte zugeordnet worden sind

MinPts = 2, epsilon = 0.5

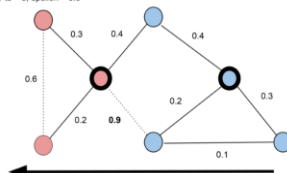
**VORTEILE:**

- Die **Anzahl Cluster** müssen **nicht vordefiniert** werden und es können **beliebig gross** sein
- Es können Cluster gefunden werden, die k-Means nicht erkennen kann (diverse Grösse und Form)
- Es ist **Rauschen gegenüber robuster** als normales Graph-based Clustering (kann **Outliers erkennen und entfernen**)
- Erfolgreich in vielen Clustering-Anwendungen

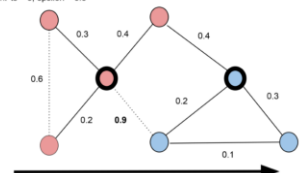
NACHTEILE:

- Zwei **Parameter müssen definiert** werden (MinPts & ϵ)
→ wenn die Distanz ϵ zu klein gewählt wird, entfallen viele Punkte als Noise
- → wenn die Distanz ϵ zu gross gewählt wird, werden mehrerer Cluster zusammengefügt
- Wenn die **Dichte variiert**, was in hoch-dimensionalen Datensätzen üblich ist, bleiben viele Cluster unentdeckt (man müsste für versch. Bereiche unterschiedliche ϵ definieren können)
- Die Resultate sind **initialisierungsabhängig**

MinPts = 3, epsilon = 0.5



MinPts = 3, epsilon = 0.5



HIERARCHISCHES CLUSTERING

MERKE: Beim Centroid- oder Graph-basierten Clustering wird der Datensatz einfach in Cluster unterteilt → Es existiert sonst keine Hierarchie → **flache Hierarchie**

In der Realität gibt es aber oft **Cluster innerhalb von Clustern** → Es existiert also eine **Hierarchie**
 → hierarchisches Clustering versucht eine Hierarchie der Cluster eines Datensatzes zu erkennen

ALGORITHMUS:

1. Jeder Punkt wird zunächst als sein eigenes Cluster definiert
2. Distanzen werden verglichen
3. Schrittweise werden die zwei nächsten Punkte/Cluster zusammengefasst
4. Sobald nur noch ein Cluster existiert, wird gestoppt

BENÖTIGTE INFORMATIONEN

Man braucht ein Ähnlichkeitsmass, um zu entscheiden, ob eine Distanz zwischen 2 Cluster als ähnlich gilt

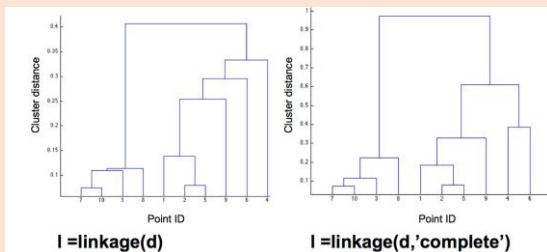
- **Single link** (minimale Distanz)
- **Average link** (durchschnittliche Distanz)
- **Complete link** (maximale Distanz)

VORTEIL:

- Ermöglicht bessere Einblicke in die Datenstruktur aufgrund der Hierarchie

NACHTEIL:

- Unterschiedliche Algorithmen produzieren unterschiedliche Clusters



- Gewünschter Output für weitere Verwendung ist oft ein flaches Clustering → unklar, wo die Hierarchie "geschnitten" werden soll

SUMMARY

- Clustering **findet Gruppen ähnlicher Objekte** in einem gegebenen Datensatz
- Die drei populärsten Familien von Clustering-Methoden:
 - o Centroid-basiertes Clustering (K-means)
 - o Graphen-basiertes Clustering (inkl. dichte-basiertes Clustering, DBSCAN)
 - o Hierarchisches Clustering
- Bei der Anwendung dieser Algorithmen ist folgendes zu beachten:
 - o Man muss sich der **Vor- und Nachteile der Algorithmen bewusst** sein
 - o Die **verwendeten Parameter** (z.B. Anzahl Cluster, verwendete Distanzfunktion) müssen **unbedingt angegeben (reported)** werden

Muster in Samples (aus OMICS-Daten) werden entweder durch **Feature Selection** oder **Clustering** gesucht/erkannt:

Feature Selektion:

- Existenz der Gruppen ist **bekannt**
- Uni-/multivariant)
- Enrichment analysis
- Classification

Clustering:

- Existenz der Gruppen ist **unbekannt**
- Visuelle Inspektion
- PCA

ANWENDUNGEN VON CLUSTERING UND DIMENSIONSREDUKTION

Bei der Analyse von Omics-Daten und Feature Selection ging es um Data-Mining mit vorhandenem Wissen über verschiedene Gruppen und Labels. Beim Clustering ist dieses Wissen nicht vorhanden.

- Verwenden von Omics-Daten als Indikator für den dynamischen zellulären Zustand
- Clustering wird auf bedingungsabhängige Omics-Daten angewendet, um Ähnlichkeiten der Zustände zwischen Proben/Samples (verschiedene Genotypen, Bedingungen etc.) ausfindig zu machen und daraus auf Ähnlichkeiten der intrazellulären Regulation zu schließen

ANWENDUNGEN VON CLUSTERING

Clustering aggregiert Samples/Features auf der Basis von paarweisen Unähnlichkeiten zwischen den Samples/Features.

METRICS:

- Unähnlichkeiten:
 - o Euklidische Distanz (direkter Abstand)
 - o Manhattan (Distanz entspricht der Summe der absoluten Differenzen der Einzelkoordinaten)
- Ähnlichkeiten:
 - o Korrelation (Pearson, Spearman), kann z.B. mit $\sqrt{1 - R^2}$ in Unähnlichkeit umgewandelt werden

CLUSTERING-TECHNIKEN:

- Centroid-basiert → k-Means
- Graph-Clustering → DBSCAN
- Hierarchisches Clustering

MERKE: Methode, Metrik und Parameter des Clustering müssen immer angegeben werden!

BEISPIELE FÜR CLUSTERING-ANWENDUNGEN

- **Discover/classify cell types** based on functional (omics) data when it's not clear a priori how many types exist
 - Example 0 (*which time profiles are similar?*)
 - Example I (*which cancer lines are similar?*)
 - Example II (*how many types of neurons exist?*)
- **Denoise** data (i.e. in interaction studies)
 - Example III (*protein complex detection*)
 - Example IV (*protein complex detection*)
- **Functional genomics** (propose function of genes/proteins based on similarity of omics profiles)
 - Example 0 (*what gene families are expressed when?*)
 - Example V (*infer protein complexes from transcriptomics*)
 - Example VI (*infer gene function from transcriptomics*)

CLUSTERING VON SAMPLES

- Clustering von Samples über alle Features hinweg.
- Die Identifizierung erfolgt auf der Basis von funktionellen (Omics)-Daten
- Beispiele: Verschiedene Mutanten, Bedingungen oder Behandlungen

- Erlaubt die **Identifizierung von Zelltypen** oder **Ähnlichkeiten in (intrazellulären) dynamischen "Antworten"**
- *Beispiele: Identifizierung des zeitlichen Ablaufs von Events, von Reaktionen auf Wirkstoffe, von Protein-Komplexen als phänotypische Antwort auf Knock-outs.*

CLUSTERING VON FEATURES

Clustering von Features über Samples hinweg

- Beispiele: Gene, Metaboliten, Proteine
- Erlaubt die Identifizierung korrelierender/coexprimierter Features
- Wird oft mit einer Enrichment Analyse (mit GO-Terms) kombiniert
- *Beispiele: Wann werden welche Gen-Familien exprimiert? Ableiten von Protein-Komplexen von Transcriptomics*

CO-CLUSTERING

- Gleichzeitiges Clustering von Samples und Features
- Dient der Visualisierung, dem **Denosing** und der Vereinigung

DIMENSIONSREDUKTION

- Nicht alle Omics-Daten sind informativ oder unabhängig
- Ziel: Dimensionen einer mehrdimensionalen Matrix (many features = many dimensions) **reduzieren**, dabei aber die **eigentlichen Informationen zu erhalten** → **Redundanzen** und **Rauschen entfernen**
- unterscheidet sich von der Feature Selection → man hat vor der Analyse keine Information über Gruppen von z.B. Samples → es wird nicht auf bekanntem Wissen aufgebaut
- Während der **Komprimierung** der Daten werden die originalen Dimensionen **transformiert/kombiniert**, sodass sie die wichtigsten Informationen in neuen Dimensionen reflektieren
 - o Dimensionen entsprechen den Features
- Es gibt viele Methoden zur Dimensionsreduktion → hier wird aber nur **PCA** besprochen

PCA

WARUM PCA? → DIMENSIONSREDUKTION!

Eine der am häufigsten verwendeten Methoden, um **Dimensionen** (Variablen) eines **multidimensionalen Datensatzes zu reduzieren (Kompression)** (aber trotzdem den grössten Teil der Varianz beizubehalten)

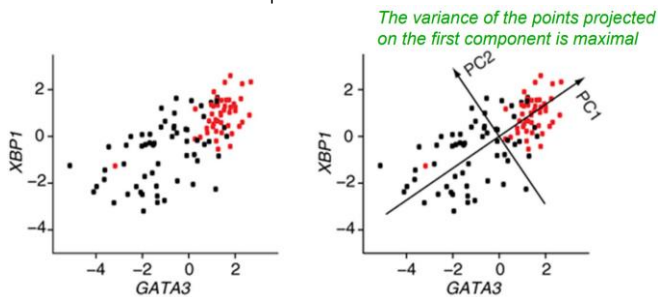
→ Bestimmung von **Principal Components (PCs)** in mehrdimensionalen Daten

(PCA can be used to explore data sets in which thousands of variables have been measured → By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables → Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped)

ZIELE

- **Dimensionsreduktion (Redundanzen reduzieren)**
- Visualisierung (Mustererkennung- Klassifizierung)
- Qualitätskontrolle
- Vorverarbeitung (**Denosing**) von Daten vor der Anwendung von Data-Mining Methoden
- Fehlende Daten abschätzen

- PCA is designed to identify directions with the largest variation and not directions relevant for separating classes of samples!
- PCA can serve as a useful first step before clustering or classification of samples.



- Jede Komponente ist durch einen Vektor (PC) definiert, der eine **lineare Kombination der ursprünglichen Merkmale** ist → Die Vektorelemente stellen die Gewichte/ Belastungen für jedes der ursprünglichen Merkmale dar

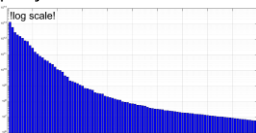
PCs = Vektor

- o sind **lineare, orthonormale** (linear unabhängig und Länge = 1) Kombinationen der ursprünglichen Dimensionen (=Features)
- o PCs werden sequenziell bestimmt, und zwar so, dass die **Varianz in Richtung der PCA maximal** ist
- o PCs nehmen in ihrer Wichtigkeit ab, tragen also zunehmend weniger Informationen
- o Die Richtung des PCs ist egal
- o Es ist wichtig sicherzustellen, dass die ersten PCs nicht durch nicht-biologische Parameter dominiert werden (z.B. Wochentag, Mitarbeiter, Apparatur etc.)

- **Loadings:** Richtungen der PC-Basis bezüglich der Standard-Basis

- **Scores:** Die transformierten Datenpunkte (also jene bezüglich der PC-Basis)

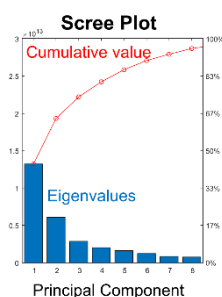
- **Eigenwerte:** Varianz der auf jede Hauptkomponente projizierten Punkte → nimmt per Definition stetig ab



- Daten werden immer zentriert (Mittelpunkt der Daten wird in den Ursprung der Standard-Basis gelegt)
- Daten können vorher skaliert werden

ANZAHL PCS

- Es werden so viele PCs (k) behalten (**oft 2**), dass ein bestimmter Anteil der Varianz (**z.B. 75%**) erklärt wird
- Man wählt all jene PCs, deren Eigenwerte grösser sind als der Durchschnitt der Eigenwerte
- Man wählt all jene PCs, bis im Screen-Plot (Auftragung der Varianz bzw. der kumulierten Varianz) der letzte "Knick" erfolgt



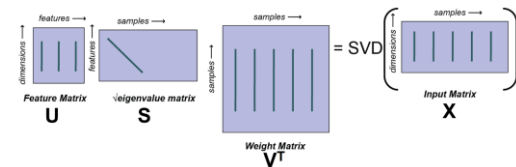
MATLAB

In Matlab gibt man als Input eine Matrix X, die in den **Spalten die Samples** und in den **Zeilen die Dimensionen** enthält

In Matlab:

```
>> [components, projections, variances] = pca(X)
```

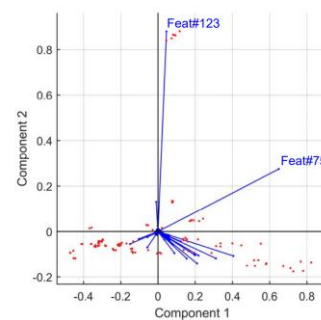
... uses Singular Value Decomposition (after centering of X to 0)
 $[U, S, V] = \text{SVD}(X) \quad \Rightarrow \quad X = USV^T$



BI-PLOT (MEIST PC1 UND PC2)

In einem Bi-Plot werden die **Sample Scores** (Datenpunkte in der PC-Basis) und die **Feature Loadings** (die ursprünglichen Richtungen der Vektoren) skaliert und übereinandergelegt

→ So können Assoziation festgestellt und betont werden

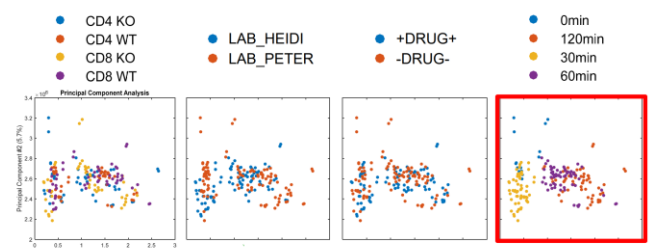


- **Samples**, die nahe beieinander sind, haben ein ähnliches Profil
- **Features**, die nahe beieinander sind, sind korreliert
- **Samples**, die nahe bei einem **Feature** sind, haben hohe Levels desselben Features
- **Samples**, die gegenüber einem **Feature** sind, haben tiefe Levels desselben Feature

VISUALISIERUNG

Man kann PCs mit Vorinformationen überlagern, um zu sehen, ob ein gewisses Muster vorhanden ist

PCs don't change. We just use a different coloring scheme to visually inspect whether a specific pattern is present



ZUSÄTZLICHE ERKLÄRUNG DER PCA:

- Die Hauptkomponentenanalyse (PCA) wird häufig in der **Clusteranalyse** und zur **Reduzierung der Dimension** des Parameterraums verwendet, insbesondere dann, wenn man noch **keinerlei Vorstellung** (Modell) von der Struktur der Daten hat
- Die Grundannahme für die Verwendung der PCA zur Clusteranalyse und Dimensionsreduktion lautet:
Die Richtungen mit der größten Streuung (Varianz) beinhalten die meiste Information
- PCA zur Reduzierung der Dimension und zur Clusteranalyse einzusetzen ist aber nicht immer möglich
→ Ob die Grundannahme, dass die Richtungen der grössten Streuung auch wirklich die interessantesten sind, zutrifft oder nicht, hängt vom jeweils gegebenen Datensatz ab und lässt sich oft nicht überprüfen (Ist gut, wenn man neben den ersten beiden auch andere PCs miteinander vergleicht!)

SUMMARY

- Dimensionsreduktion wird (auch) durchgeführt, um Daten zu **visualisieren** und von der menschlichen Fähigkeit, Muster gut zu erkennen, Gebrauch zu machen
- Dimensionsreduktion macht per se kein Clustering
- PCA ist **eine von vielen Methoden**
- PCA **bewahrt die Varianz** und ist gut darin, systematische Fehler zu erkennen (Qualitätskontrolle)
- Werden durch PCA keine Cluster ausfindig gemacht, bedeutet das nicht, dass keine univariat/multivariat unterschiedliche Features existieren um zwischen bekannten Gruppen zu unterscheiden → **Signifikant unterschiedliche Features beeinflussen die PCA nur, wenn ihre Werte genügend gross sind, um die Gesamtvarianz zu beeinflussen!**

MERKE:

Wenn a priori **keine Gruppen bekannt** sind, benutzt man Clustering (PCA) oder Dimensionsreduktion

Wenn a priori **Gruppen bekannt** sind, verwendet man univariate/multivariate Feature Selection, Enrichment Analysis oder Klassifizierung

LINK PREDICTION

Mithilfe von Link Prediction will man in einem gegebenen Netzwerk **fehlende Verbindungen vorhersagen**

- **Interaktionsvorhersage** → Durch das Wissen von Bekannten Interaktionen wollen wir herausfinden, welche Interaktionen noch wahrscheinlich sind
- **Wird verwendet, um**
 - o Genregulationsnetzwerke zu entdecken
 - o Gen Co-Expression zu finden
 - o Protein-Protein Interaktionen vorauszusagen
- Ist auch ausserhalb der Biologie weit verbreitet
- Um Link Prediction durchzuführen, verwendet man oft **diverse Datentypen**
- Arten von Informationen, die in Link-Prediction miteinfließen (Link prediction starts with):
 - o Wissen über Expression der jeweiligen Gene
 - o Bekannte (aber inkomplette) Interaktionsnetzwerk Struktur
 - o Gen-/Protein Sequenzinformationen etc.

BEISPIELE:

- Man will ein Genregulationsnetzwerk ergänzen. Welcher Transkriptionsfaktor reguliert welches Gen? Welche Gene werden koexprimiert?
- Man hat ein Netzwerk von Protein-Protein Wechselwirkungen und will fehlende Interaktionen ergänzen
- Netflix: Neue Filme vorschlagen

ZWEI TYPEN VON LINK PREDICTION:

1) Unsupervised Link Prediction:

2) Supervised Link Prediction:

1. UNSUPERVISED LINK PREDICTION:

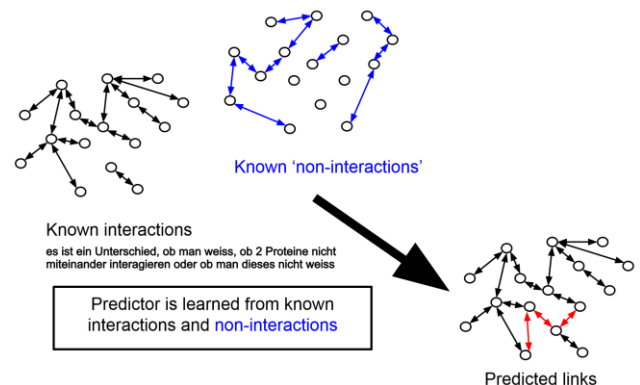
- Das Modell basiert nicht auf dem Lernen von Beispielen, sondern auf **vordefinierten Regeln** (z.B. Definition einer Interaktion)
- Vorhersage beruht auf einer Regel, welche nicht aus den Daten/Beispielen gewonnen wurde

SIMILARITY-BASED APPROACH

- Eine Verbindung zwischen Genen a und b wird vorhergesagt, wenn die **Ähnlichkeit $s(a,b)$** über einem **Grenzwert θ** ist
- Kriterien für die Ähnlichkeits-Definition ($s(a,b)$)
 - o Pearson's Correlation Coefficient
 - o Gemeinsame (mutual) Information
 - o String Kernels, die gemeinsame Teilsequenzen in zwei Proteinen (k-mere) zählen
 - o Anzahl gemeinsamer Nachbarn
- **Vorteil:**
 - o Einfach zu implementieren/ programmieren
 - o Skalierbar auf grosse Netzwerke.
- **Nachteile:**
 - o Wie soll θ definiert werden?
 - o Ähnlichkeit ist nicht immer eine Bedingung für Interaktion!

2. SUPERVISED LINK PREDICTION:

- Es sind Beispiele von existierenden Verbindungen gegeben und das Modell kann **von diesen Beispielen für Vorhersagen lernen**
- Algorithmus sucht Eigenschaften in den Daten, um eine eigene Regel zu formulieren



- Es gibt diverse Ansätze:
 - o 2.1 cluster based learning
 - o 2.2 latent group models
 - o 2.3 latent feature models
 - o 2.4 similarity-based classification
 - o 2.5 Regressions-basierte Ansätze

2.1 CLUSTER-BASED LEARNING

Eine Verbindung zwischen a und b wird vorhergesagt, wenn **a und b zu dem gleichen Cluster gehören**

- Die Cluster-Zugehörigkeit entspricht einer latenten Variablen → Interaktionswahrscheinlichkeit hängt von 1 Variablen pro Punkt ab (die ihre Clusterzugehörigkeit bestimmt)
- Vorteil:
 - o Allgemeiner als Verbindungen aufgrund von paarweisen Ähnlichkeiten vorherzusagen
- Nachteil:
 - o Biologisch ist es eventuell unrealistisch, dass Interaktionen nur von der Cluster-Zugehörigkeit abhängen (one latent variable)

2.2 LATENT GROUP MODELS

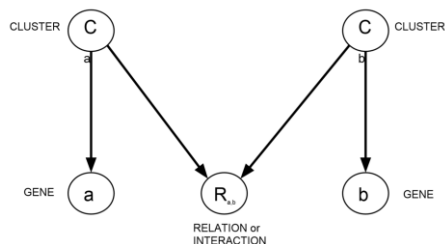
- Wahrscheinlichkeiten für Interaktionen zwischen Genen werden **basierend auf Cluster-Zugehörigkeiten (latente Gruppen)** hergeleitet
- Die Methode wird in eine Trainings- und eine Vorhersage-Phase unterteilt

Trainings-Phase:

1. Eine **Untergruppe S von Genen** wird ausgewählt
2. Die Gene in S werden basierend auf den Genexpressionsprofilen **in k Cluster** unterteilt
3. Für jedes Cluster-Paar i und j wird eine empirische **Interaktionswahrscheinlichkeit p_{ij}** bestimmt

Vorhersage-Phase:

1. Gegeben sei ein neues Paar von Genen a und b
2. a und b werden den ähnlichsten Clustern C_a und C_b zugeordnet
3. Die Interaktionswahrscheinlichkeit entspricht p_{ij}

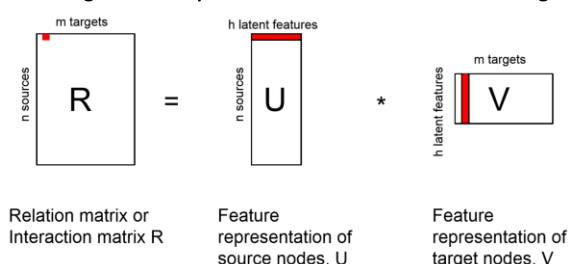


2.3 LATENT FEATURE MODELS

Bei der Latent Feature-Based Link Prediction basierend die Vorhersagen auf **mehreren latenten oder versteckten Features (mehrere latente Variablen)**

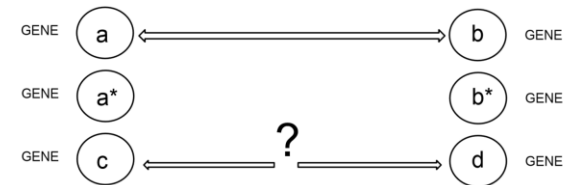
→ höhere dimensionale Beschreibung, die sinnvoller für Biologen ist

Das Vorgehen entspricht einer Matrix-Faktorisierung:



- **Matrix R:** repräsentiert die **Interaktionen** → sind nicht alle bekannt! $R = U \cdot V$
- **U:** besteht aus den **Sources** (z.B. Transkriptionsfaktoren) der h latenten Features,
- **V:** besteht aus den **Targets** (z.B. den Genen) der h latenten Features
- Linkvorhersage bedeutet hier Matrixvervollständigung

2.4 SIMILARITY-BASED CLASSIFICATION (KERNEL-APPROACH)



Wir wissen:

- Paare von Genen (a,b) interagieren miteinander
- Paare von Genen (a*,b*) interagieren nicht miteinander
- Wissen wir nicht/ wollen wir herausfinden: Interagieren neue Paare von Genen (c,d) miteinander?

Genereller Ansatz: Man bestimmt über ein **Ähnlichkeitsmass**, ob (c,d) ähnlicher zu den bekannten Interaktionen (a,b) oder zu den bekannten Nicht-Interaktionen (a*,b*) ist

→ Unbekannte Interaktion wird mit existierenden Interaktionen und Nicht-existierenden Interaktionen verglichen und als jene definiert, deren sie ähnlicher ist

ÄHNLICHKEITSMESSUNG Z.B. VIA KERNEL-FUNKTION

Tensor Pairwise Kernel

Gegeben sind zwei Paare (a,b) und (c,d):

$$k_{\text{tensor}}((a,b), (c,d)) = k_n(a,c) \cdot k_n(b,d) + k_n(a,d) \cdot k_n(b,c)$$

k_n : Kernel, der die Ähnlichkeit zweier Knoten misst.

Metric Learning Pairwise Kernel

Gegeben sind zwei Paare (a,b) und (c,d):

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

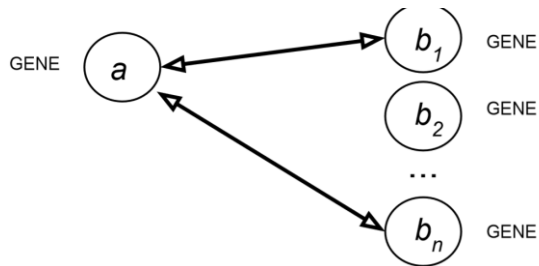
$\varphi(g)$ ist ein Vektor, der die Features von Gen/Protein g beschreibt.

Ein Paar (a,b) ist ähnlich zu einem Paar (c,d), wenn $a - b$ ähnlich zu $c - d$ ist, also die Unterschiede ähnlich sind.

PITFALL: AUSWAHL DER NEGATIV BEISPIELE:

- gute Auswahl von Negativbeispielen (= Nicht-Interaktionen) zu treffen ist **schwierig**
- Allgemeine Strategie: Man wählt Proteine von verschiedenen Zellkompartimenten gemäss 'Gene Ontology (GO) Cellular Component Annotation'
 - o Negativ Beispiele aus anderen Zellkompartimenten zu wählen 'oversimplifies' und ist zu optimistisch
- andere Möglichkeit: Predictors für Datensätze zu evaluieren, die die gleiche Anzahl positive und negative Beispiele für Interaktionen haben
 - o ist zu pessimistisch
- Weitere Strategie: Generierung eines zufälligen negativen Beispiels

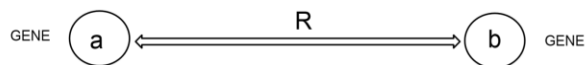
2.5 REGRESSIONS-BASIERTE ANSÄTZE



Man will den Zustand eines Gens basierend auf dem Wissen, das man über alle anderen Gene hat, voraussagen → mittels **LASSO**

- Jedes Gen wird durch einen Vektor repräsentiert
 - Um Interaktionspartner von Gen a zu finden, führt man eine **Regression** zwischen a und allen anderen Genen $\{b_1, \dots, b_n\}$ durch
 - Interaktion für Gene, die im Regressionsmodell eine Gewichtung von **ungleich Null** erhalten
 - **LASSO** = 'linear regression with sparsity constraints' → berücksichtigt unsere Annahme, dass ein Gen nur mit wenigen anderen Genen wechselwirken sollte
- ➔ **Schwierigkeit:** Wie werden die Regularisierungsparameter, die die Seltenheit der Interaktionen belohnen, gewählt? → L1 sollte muss werden, um die Seltenheit einer Lösung kontrollieren zu können

AUSWERTUNG DER LINK-PREDICTION



True Label (R)

Predicted Label (R*)

	R = 1 (true-P)	R = -1 (true-N)
R* = 1 (Predicted-P)	TP (true positive)	FP (False-positive)
R* = -1 (Predicted-N)	FN (False-negative)	TN (True-negative)
	P	N

EVALUATIONSKRITERIEN:

- **Genauigkeit:** $\frac{TP+TN}{TP+FP+FN+TN} = \frac{TP+TN}{P+N}$
= Anteil der Vorhersagen an, die korrekt sind
- **Sensitivität (Recall, True-positive-Rate):** $\frac{TP}{P}$
(Wie viele echte Interaktionen finde ich?)
- **Präzision:** $\frac{TP}{TP+FP}$ (Anteil der vorhergesagten Interaktionen (TP+FP), die korrekt sind)
- **Spezifität:** $\frac{TN}{N}$ (Wie viele der N wurden erkannt?)
- **False-positive-Rate:** $\frac{FP}{N} = 1 - \text{Spezifität}$ (Wie viele der N werden nicht erkannt?)

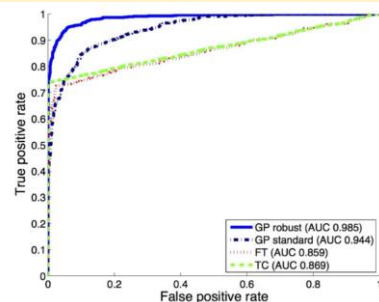
EVALUATIONSKRITERIUM ROC (RECEIVER OPERATING CHARACTERISTIC)

Resultate der Link prediction werden hauptsächlich mittels eines **ROC-plots** ausgewertet:

ROC= Receiver Operating Characteristic

- ROC plottet die **True-positive-Rate** gegen die **False-positive Rate**

- AUC = Area under the ROC curve → Man misst die Fläche unter der Kurve → = Vorhersagegenauigkeit → ist die Wahrscheinlichkeit, dass der Klassifikator richtig erkennt, welches das positive Beispiel ist
- **OPTIMAL:** die Fläche unter der Kurve ergibt 1 (**AUC = 1**) → Die bedeutet, dass die **true positive Rate maximal** ist und keine false positive Werte auftreten



PITFALLS

- Auch ein hoher AUC-Wert kann in der Praxis nutzlos sein → wenn eine Klasse viel grösser ist, als die andere

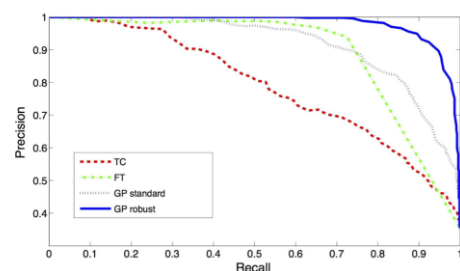
LÖSUNG:

Wenn die Klassen unterschiedlich gross sind (positive class und negative class sind stark unbalanciert), sollte man neben dem AUC-Wert immer noch die die beiden Werte

Präzision und Recall angeben:

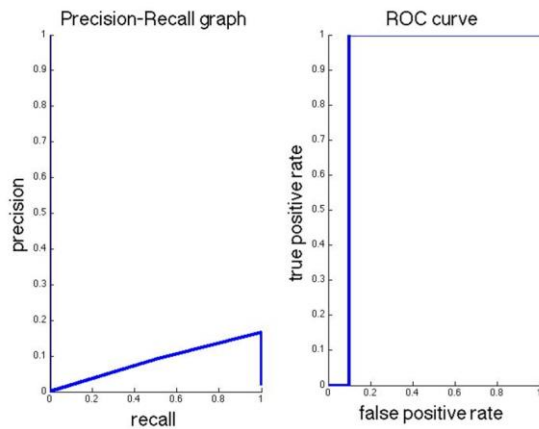
- **Präzision:** $\frac{TP}{TP+FP}$
→ Anteil der vorhergesagten Interaktionen (positive), die echte Interaktionen (true positive) sind
 - **Recall (True-Positive-Rate):** $\frac{TP}{P}$
→ Anteil aller vorhandenen Interaktionen (positive), die der Classifier finden konnte (TP)
- ➔ Aus diesen Werten kann man einen Plot erstellen (Precision-Recall-Curve)

Precision-Recall-Curve



BEISPIEL:

Beim folgenden Experiment wurde Link Prediction auf 102 Knotenpaare angewendet, wobei nur zwischen zwei Knoten tatsächlich eine Interaktion vorhanden war. Das Resultat der AUC/ROC ist irreführend und zeigt viel bessere Ergebnisse als die Precision-Recall Kurve (AUC-Wert ist viel zu gross). Es wird empfohlen, jeweils **beide Werte** anzugeben.



Rechts: AOC ist $0.9 \cdot 1 = 90\%$ Präzision → nach 10 wurden alle TP erkannt

Links: Präzision steigt (zuerst 1 von 10 dann 2 von 12) sinkt dann aber wieder → AOC/ROC sehr irreführend

SUMMARY

Biologische Netzwerke sind **von Natur aus unvollständig**.

Die Link-Vorhersage versucht, diese Netzwerke **durch rechnerische Vorhersagen zu vervollständigen**.

- Es gibt diverse Methoden
- Es ist wichtig, ihre zugrunde liegenden Annahmen zu verstehen, da die von ihnen vorhergesagten Zusammenhänge diese Annahmen widerspiegeln
- Verschiedene Strategien zur Auswahl von Negativbeispielen für die überwachte Linkprognose können zu verzerrten Ergebnissen führen
- Die Auswertung der Linkprognose wird dadurch erschwert, dass die **positive Klasse (Interaktionen) und die negative Klasse (Nicht-Interaktionen) sehr unausgewogen sind**. Es wird empfohlen, neben der AUC auch die **Genauigkeit der Berichterstattung und des Rückrufs zu berücksichtigen**

BIOLOGISCHE NETZWERKE

Kartierung biologischer Netzwerke:

- Was ist die Basis für Netzwerk-Datenbanken?
- Welche Rolle spielt die Link-Vorhersage?
- Was sind die Unzulänglichkeiten von Netzwerk-datenbanken?

"Das Netzwerk als Ziel"

Einsatz von (vorhergesagten) biologischen Netzwerken:

"Das Netzwerk als Werkzeug"

DATENBANKEN

- Experimentell verifizierte, aber auch vorhergesagte biologische Netzwerke werden **in Datenbanken** zur Verfügung gestellt
 - o Beispiele sind STRING, Reactome, TRED
- Verschiedene Datenbanken liefert verschiedene Resultate → sollten **verifiziert** werden
- **Schwierig experimentell zu verifizieren** (ist nur für wenige Interaktionen möglich)
 - o Limitierte Ressourcen (Zeit, Geld, Material)
 - o Interaktionen sind dynamisch
 - o Methoden liefern unterschiedliche Ergebnisse
- Interaktionen lassen sich also nicht umfassend und experimentell erfassen → man muss darum auf **computergestützte Methoden zurückgreifen**
 - o Netzwerke rechnerisch erweitern und verfestigen/verifizieren → Link Prediction
 - o Bei einem Netzwerk mit bekannten Interaktionen jene finden, die aktiv und wichtig für den Phänotyp sind

PROTEIN-PROTEIN-INTERAKTIONEN (PPI)

Als Beispiel → auf alle Netzwerke übertragbar!

- Es gibt viele experimentelle Methoden um PPIs nachzuweisen, aber:
 - o geringe Übereinstimmungen zwischen den Studien
 - o Methoden sind komplementär
 - o Daten sind verrauscht oder unvollständig
 - o nur ein kleiner Teil der Bedingungen kann getestet werden

Mit computergestützten Methoden will man **experimentell gefundene Interaktionen bewerten**, die **Daten bereinigen** und **unbekannte Interaktionen vorhersagen**

Folgende Features/Daten werden für PPI-Inference gebraucht:

- Experimentelle Interaction-Assays (Y2H, MS)
- Indirekte Omics-Daten (z.B. Genexpression)
- Funktionale Annotationen (GO, MIPS)
- Sequenzdaten (Domänen)
- Strukturelle Informationen von Proteinen

Validierung basiert auf früheren gross angelegten Studien und der gezielten Analyse neuartiger Vorhersagen

EINSCHRÄNKUNGEN VON LINK-PREDICTION

Problem: alle **grossbasierten Methoden sind fehleranfällig**, wie sicher sind wir uns also, eine Vorhersage als false positive oder true negative zu bezeichnen?

- PPI-Prediction hat **viele Einschränkungen**, welche aber auch für alle Link Prediction Studien (non-PPI) gelten:
 - o Link Predicting (Vorhersage von Interaktionen) ist schwieriger, wenn **relativ unbekannte, noch nicht erforschte Proteine involviert** sind (fehlende Informationen wie Tertiärstruktur, keine GO-Terme oder Domain-Annotations, wenig bekannte Wechselwirkungen etc.)
 - o schlecht in der **Vorhersage von vorübergehenden Interaktionen** (nicht Obligaten) → bei diesen Interaktionen sind die beiden kodierenden Gene nicht hoch korreliert
 - o Wenn die Vorhersagemethoden auf Basis von experimentellen Daten trainiert werden, nehmen sie die Störungen der experimentellen Daten auf (Trainingsset)

BEST PRACTICE – NETWORK RECONSTRUCTION

- Interaktionen sind zustandsabhängig
- **Experimentelle Techniken** sind nach wie vor grundlegend, um Interaktionen zunächst in grossem Massstab abzubilden
- **Rechnerische Rekonstruktionen und Vorhersagen** sind notwendig, weil es nicht möglich ist, alle Bedingungen experimentell zu testen und zu finden
- Alle experimentellen und rechnerischen Methoden sind gut für bestimmte Aufgaben
- Mehrere Methoden verwenden! → Mehrere Methoden funktionieren besser als einzelne
- Die Analyse von 35 Methoden hat gezeigt, dass jede Familie von Methoden generell besser/schlechter in der Vorhersage von bestimmten Dingen ist → Methoden von verschiedenen Familien kombinieren

VERWENDUNG REKONSTRUIERTER NETZWERKE

- Entdeckung von Funktion und Organisation
- Netzwerk = Enzyklopädie zur **Unterstützung der Dateninterpretation**
- Differentielle Analyse von Interaktionsnetzwerken zur **Identifizierung von dynamischen oder konservierten Modulen**
- Auffinden von aktiven Modulen in Netzwerken durch Kombination von funktionellen Daten

MNI-METHODE

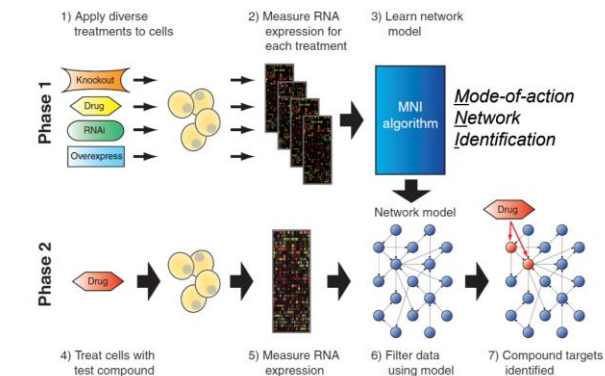


Figure 1 Overview of the MNI method. In phase 1, a set of treatments, including knockouts, compounds, overexpressions and/or RNAi, is applied to an organism. Cells or tissues are sampled, and mRNA is collected. The abundance changes of all mRNA species in the organism are measured. The data are used by the MNI algorithm to infer a model of the regulatory influences between genes in the organism (blue-filled circles indicate genes; arrows indicate regulatory influences). In phase 2, a test treatment, such as a drug, is applied to the cells and expression changes of all mRNA species are measured. The expression data are then filtered using the network model to distinguish the targets of the test treatment (red-filled circles) from secondary responders.

SUMMARY

- Interaktionen sind abhängig von den Bedingungen
- Experimentelle Methoden sind wichtig, um einen ersten large-scale Entwurf anfertigen zu können
- Computergestützte Methoden sind wichtig, weil es nicht möglich ist, alle Bedingungen zu testen
- Alle experimentelle und computerbasierte Methoden gut in bestimmten Dingen
- Mehrere Methoden kombinieren
- Netzwerke werden häufig verwendet, um physikalische Interaktionsnetzwerke (Protein-Protein, Protein-DNA,...), genetische Netzwerke (TF-Gen, Gen-Gen,...) und metabolische Netzwerke zu beschreiben.
- Netzwerke ermöglichen es, Beobachtungen in einen molekularen Kontext zu stellen, d.h. Wechselwirkungen (z.B. Synergismus, Pufferung) zu verstehen, die für ein bestimmtes Merkmal/Krankheiten/Phänotyp/Zustand/Zustand relevant sind.