Medizinische Diagnose basierend auf

DNA-Microarray Daten

In dieser Vorlesung wird:

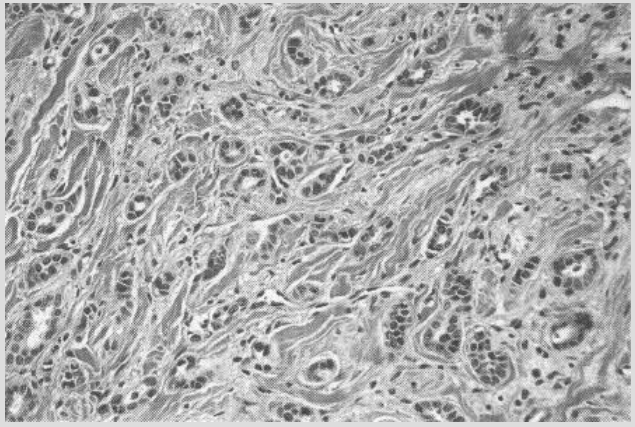
- **Das Potential einer relativ neuen Technologie (Microarrays) in der medizinischen Diagnostik dargestellt:**

Was wird wie beobachtet und wozu ist das gut ?

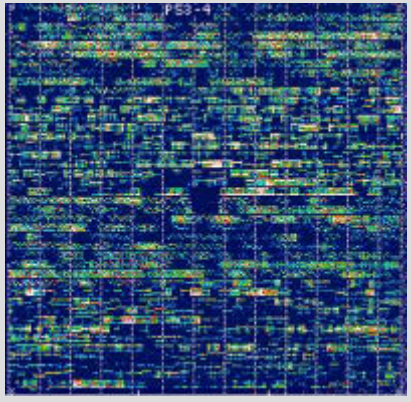
2. Ein zentrales bioinformatisches (statistisches) Problem dieses Planes umrissen

Wozu braucht man dabei Theoretiker?

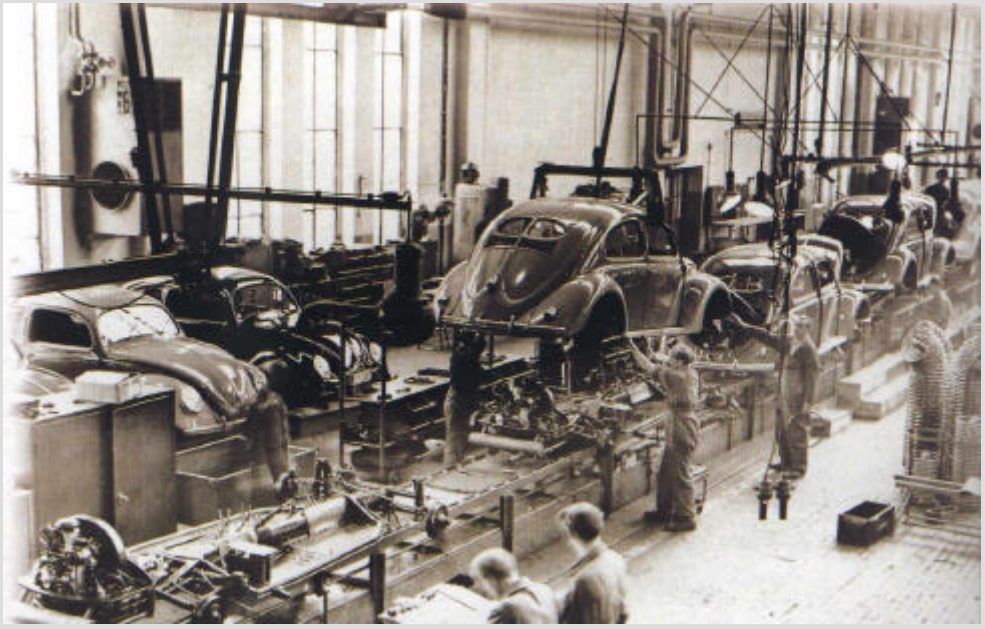
Statt des Blicks auf das Äußere der Zelle ...



... der „Blick“ ins Innere der Zelle



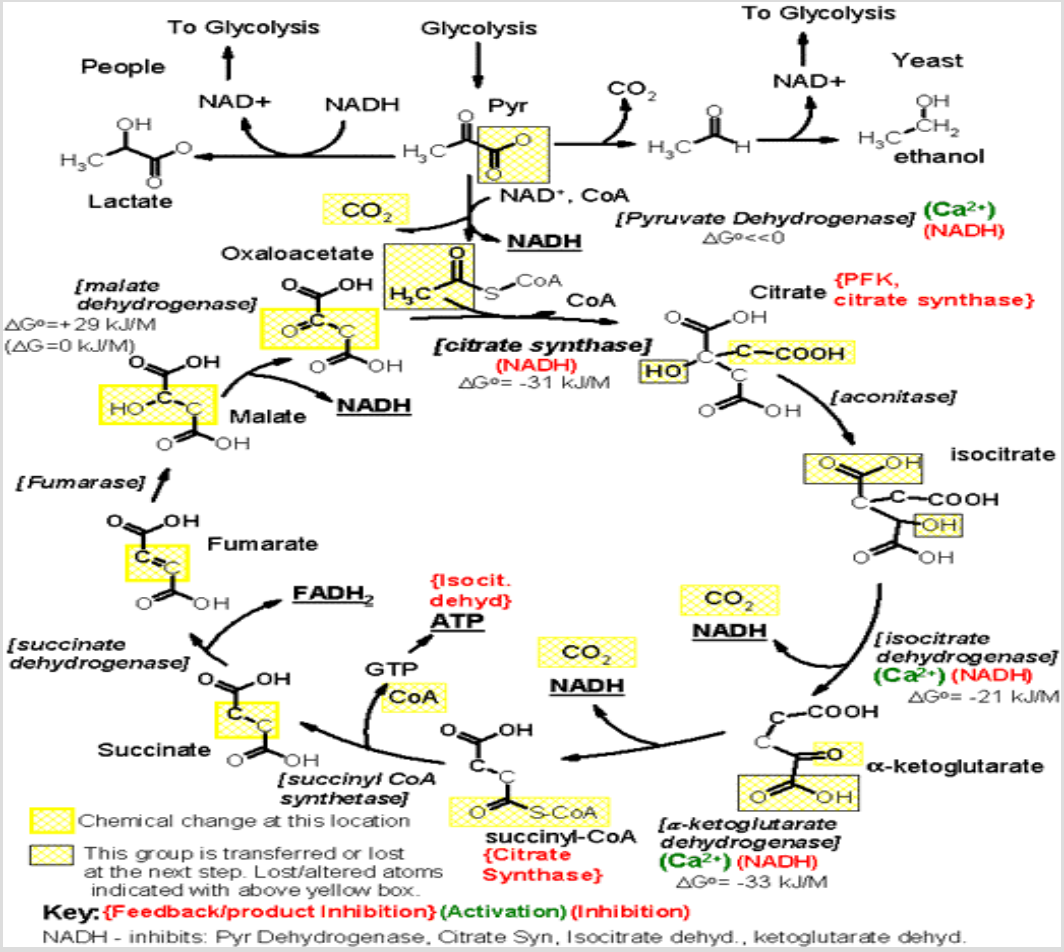
Was gibt es in Zellen zu sehen?



Action

Zellen arbeiten!

Was wird gearbeitet ?



z.B Energieproduktion

Was soll gemacht werden?

Wie soll auf besondere Ereignisse (Krankheit) reagiert werden?

Wie wird die Arbeit organisiert?

Wie soll es gemacht werden ?

Wieviel soll davon gemacht werden?

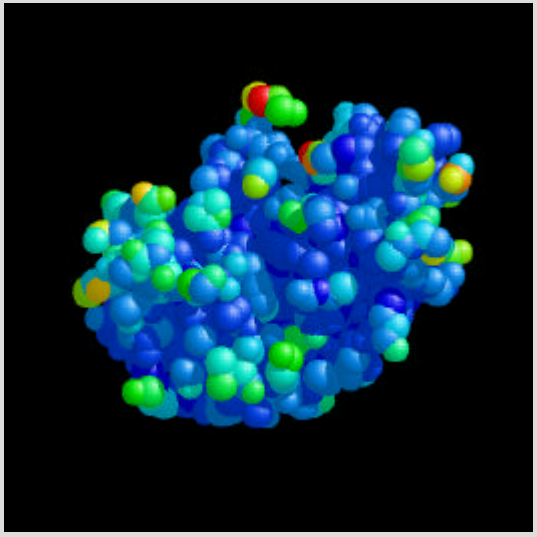
Wann soll es gemacht werden?

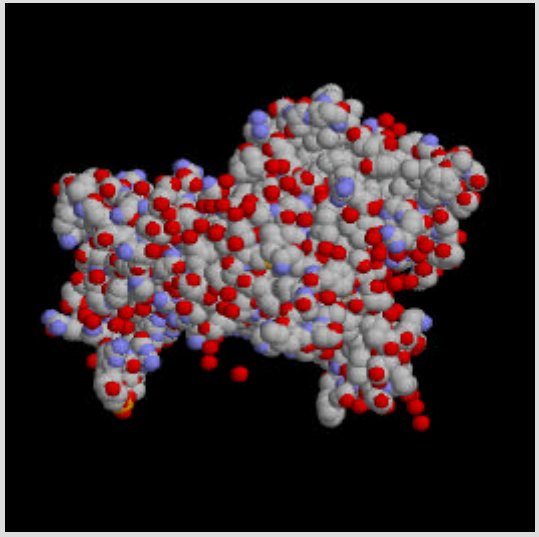
Kein Boss, sondern ....

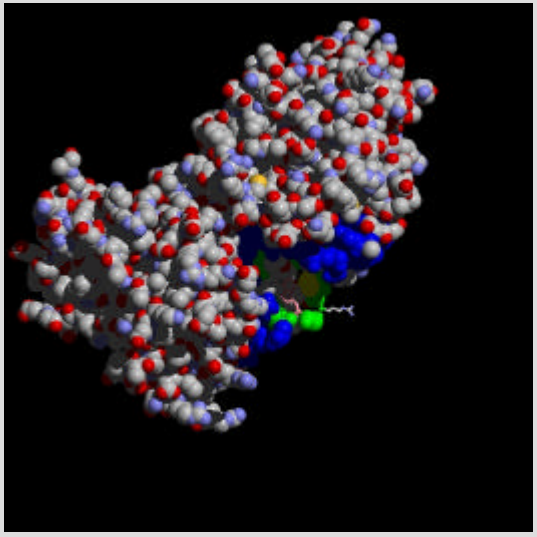
... Selbstorganisation

Wer arbeitet?

Hauptsächlich Proteine!

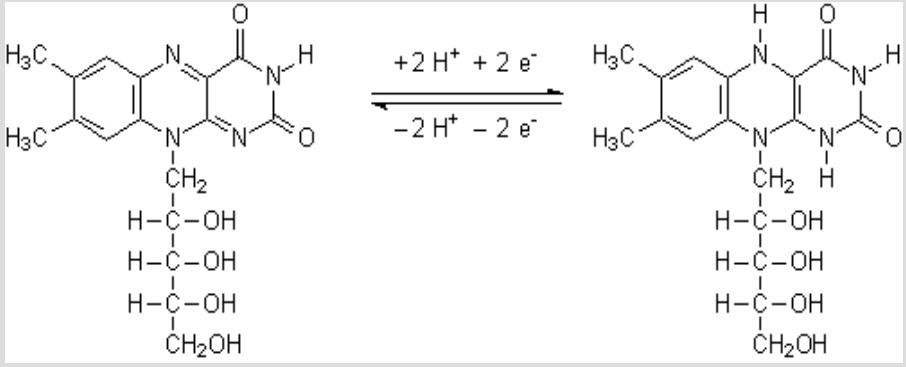


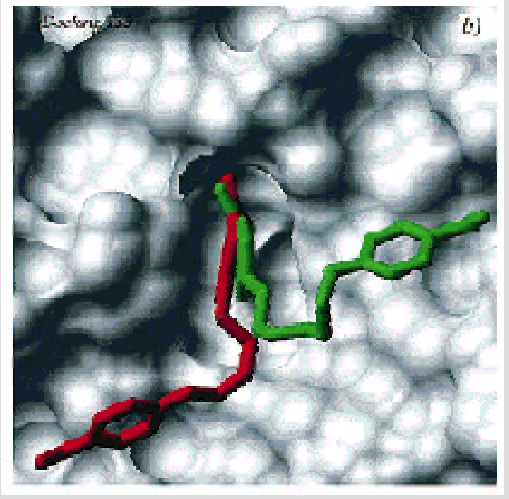




Proteine sind Facharbeiter

Die Spezialisierung ergibt sich aus ihrer 3D-Struktur

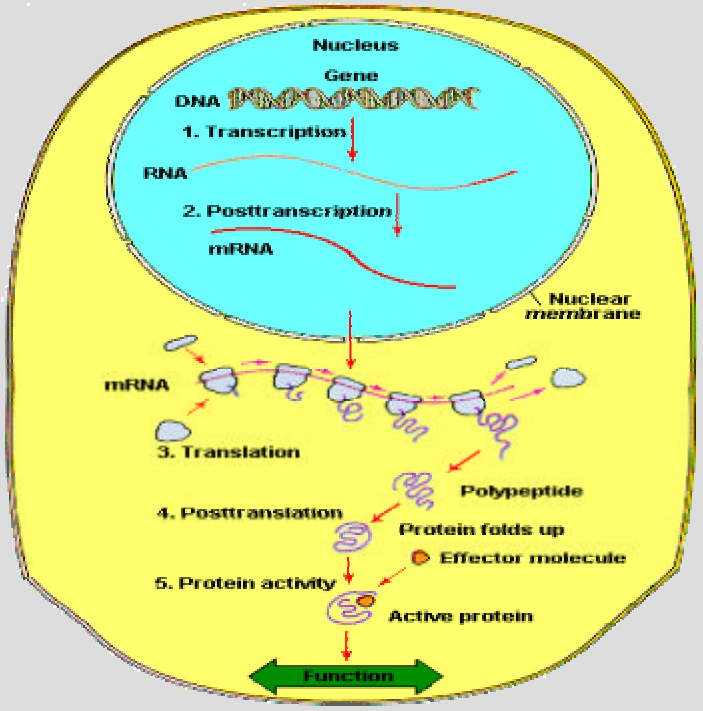




Typische Arbeitsgebiete:

Bauwesen, Chemie, Kommunikation,...

Ausbildung und Einsatz von Proteinen



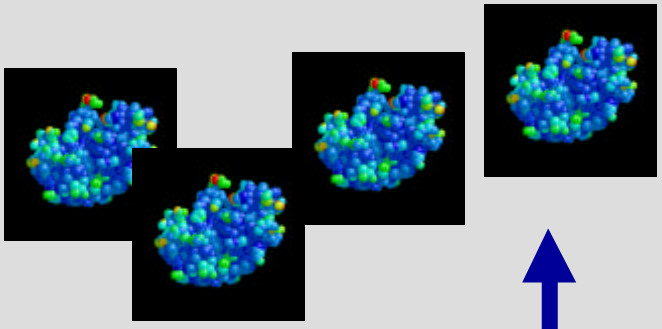
5' ...A T G G C C T G G A C T T C A... 3' Sense strand of DNA
3' ...T A C C G G A C C T G A A G T... 5' Antisense strand of DNA

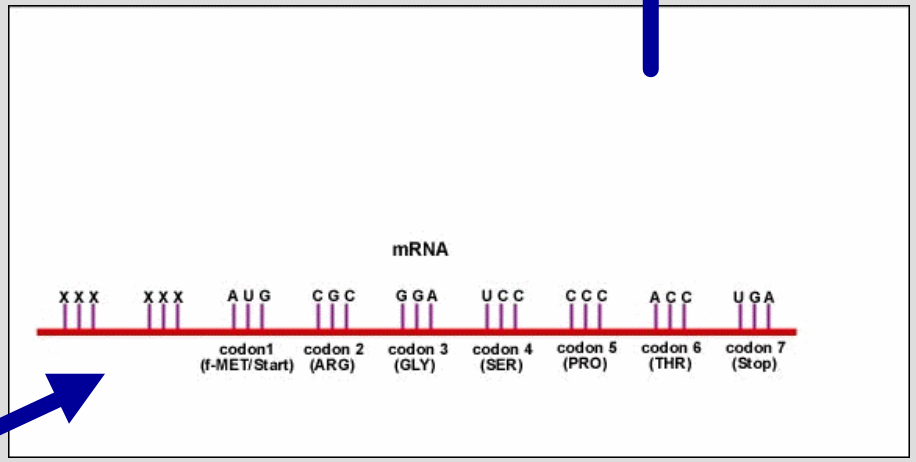
↓ Transcription of antisense strand

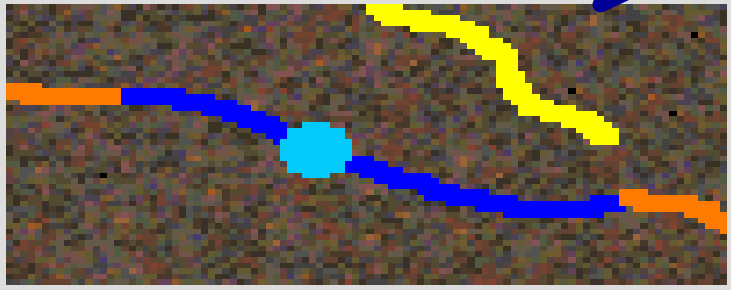
5' ...A U G G C C U G G A C U U C A... 3' mRNA

↓ Translation of mRNA

Met — Ala — Trp — Thr — Ser — Peptide

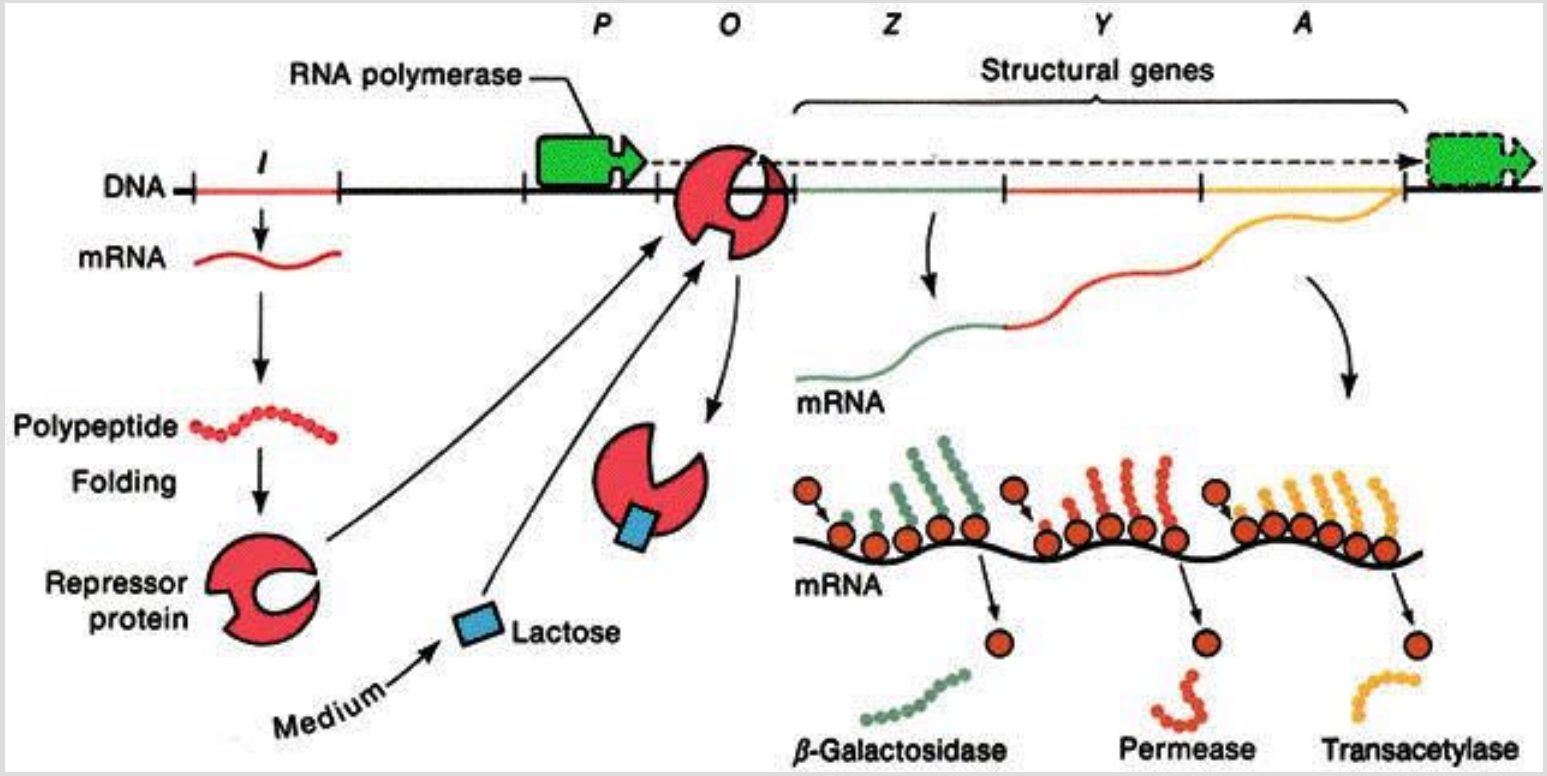






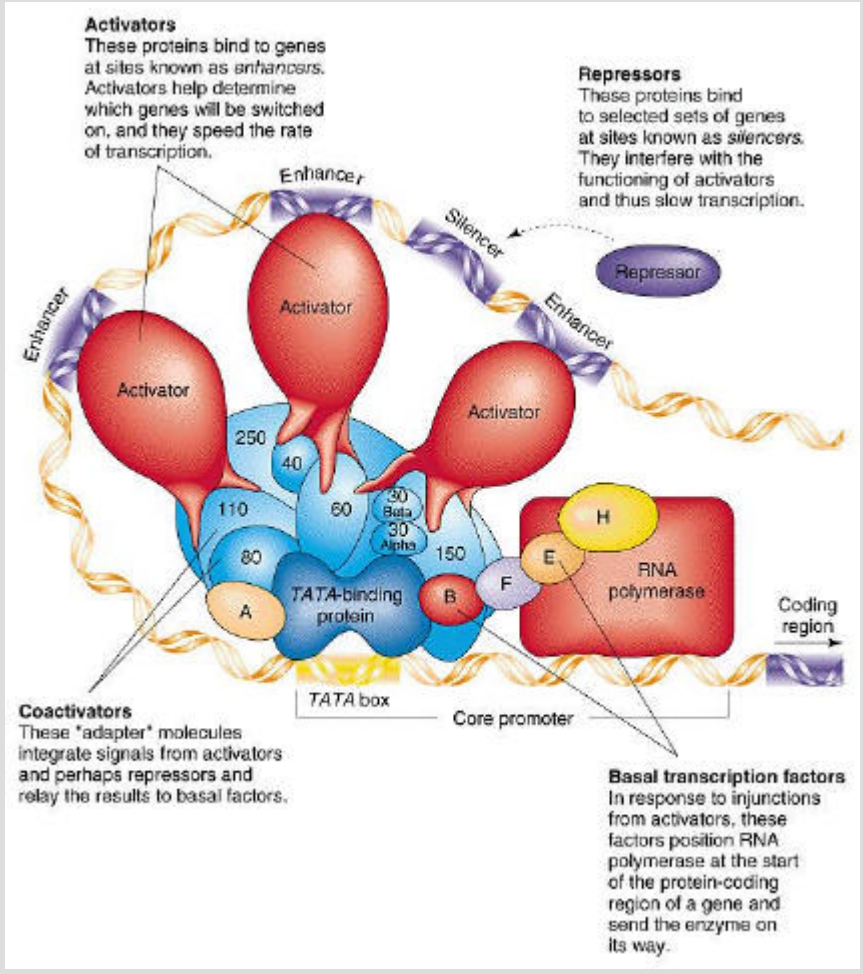
Protein Produktion

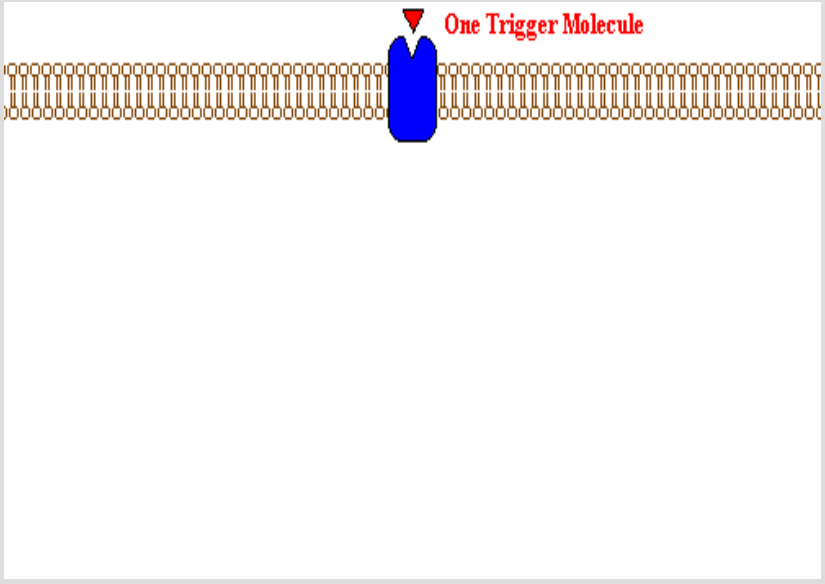
Wie wird diese Produktion gesteuert?



z.B. Lac Operon in E. Coli

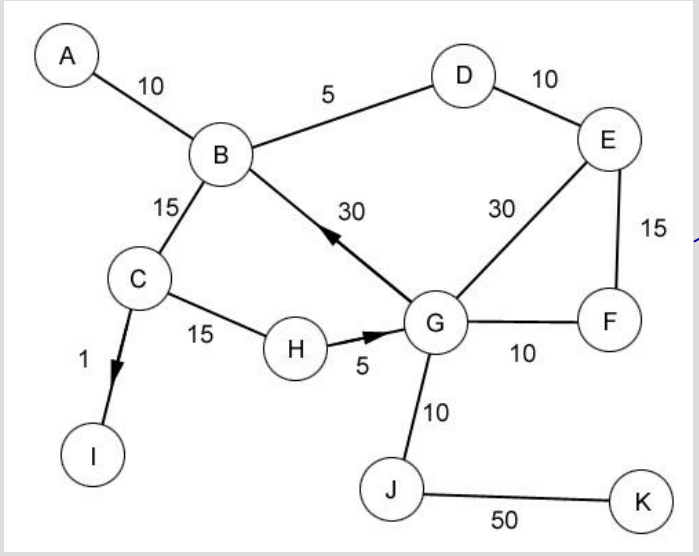
Und wie funktioniert es im Menschen?

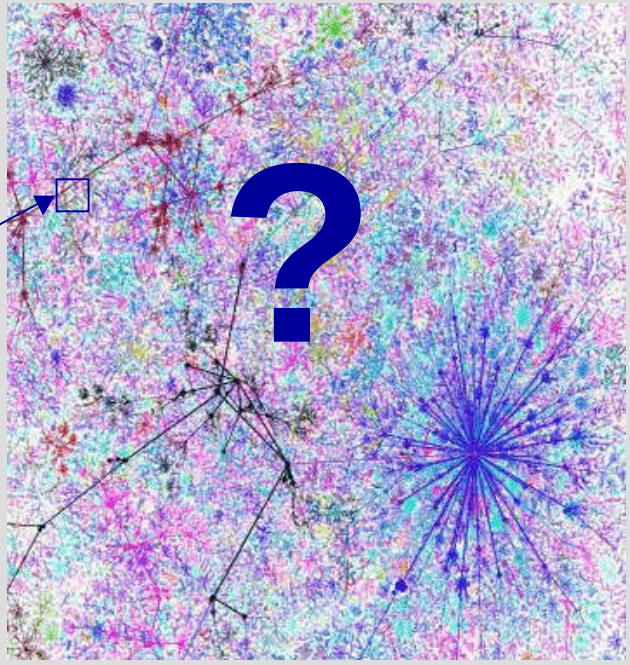




im Prinzip genauso ... nur komplizierter ...

Die Arbeit in den Zellen ist durch ein wirklich kompliziertes regulatives Netzwerk organisiert ...



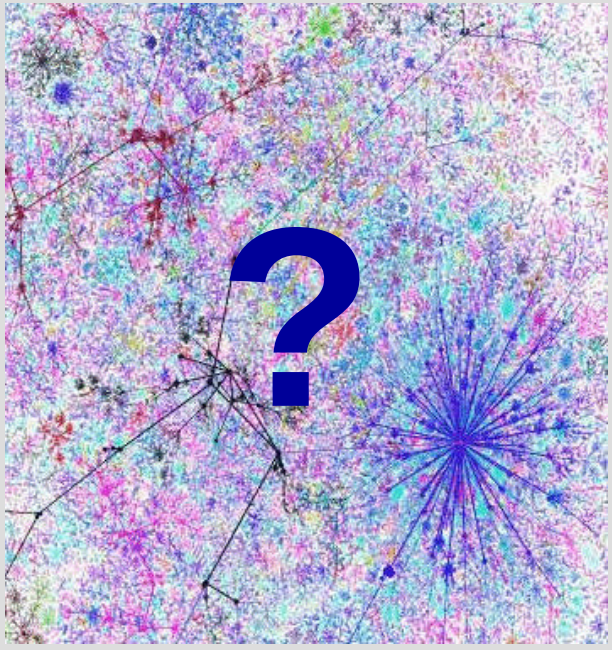


... das wir kaum kennen.

Größenordnungen:

ca. 30.000 Gene

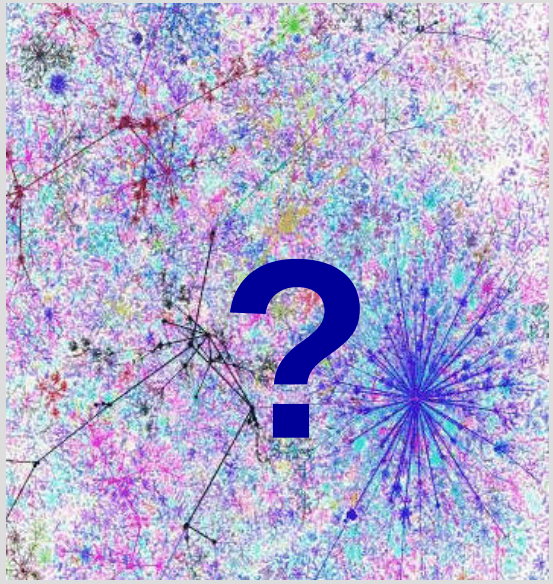
> 100.000 Transkripte



Ein wirklich sehr komplexes Netzwerk!

Dieses Netzwerk reagiert auf innere und äußere Ereignisse:

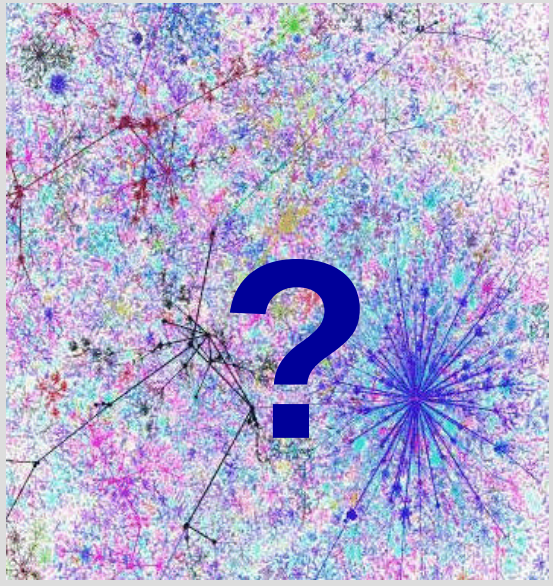
- genetische Veränderung
- Infektion
- Vergiftung
- Streß
- etc. ...



Wie?

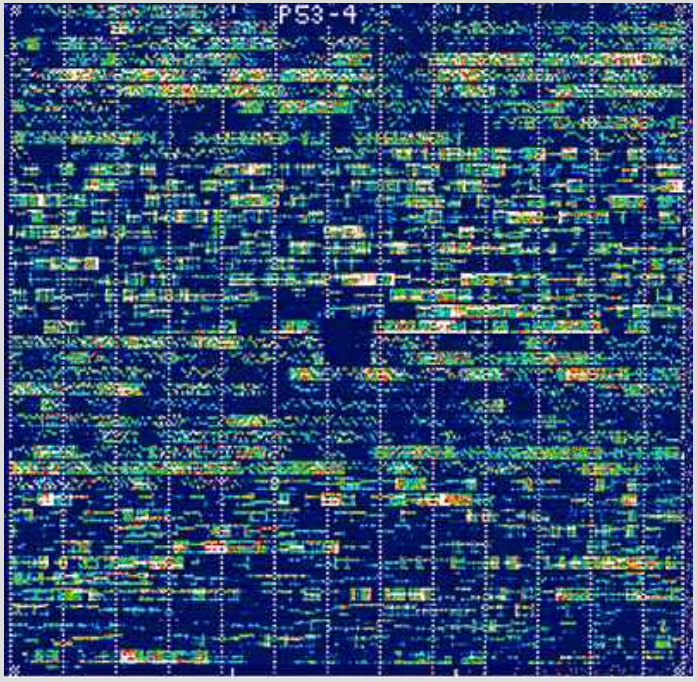
Krankheiten können (sollten) als charakteristische Zustände dieses Netzwerks verstanden (definiert) werden ...

... dazu müssen wir das Netzwerk beobachten können!



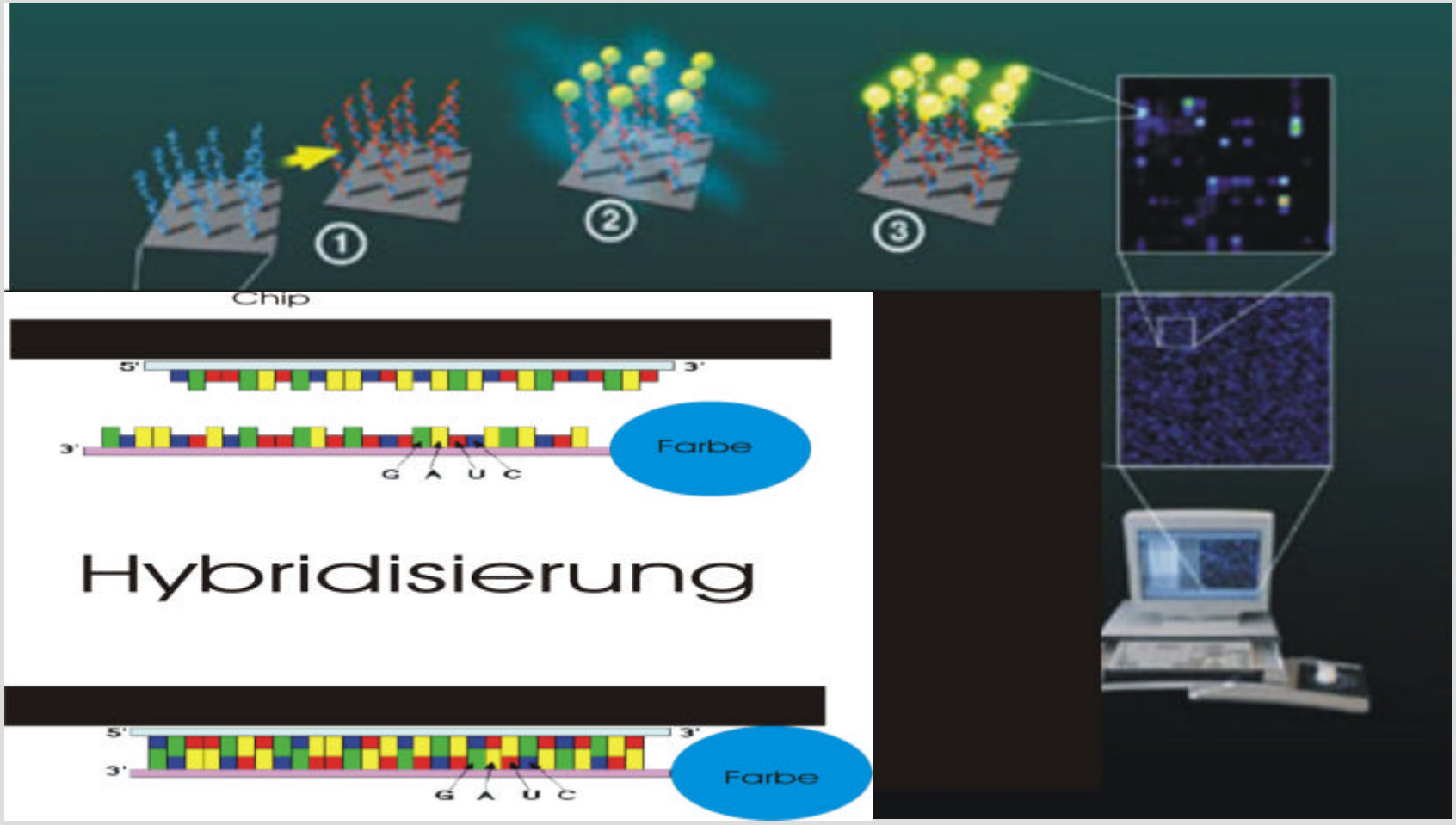
Auf einem DNA-Microarray können wir

die Häufigkeit von 1000 Transkripten (RNA Molekülen) parallel messen



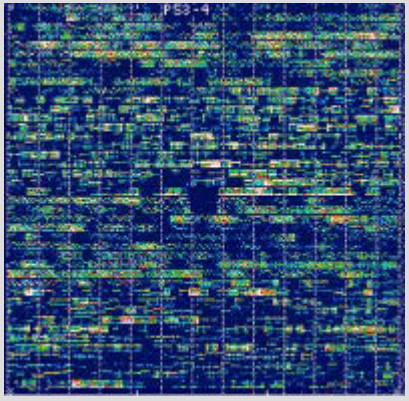
Digitales Bild der Zelle

Wie geht das?



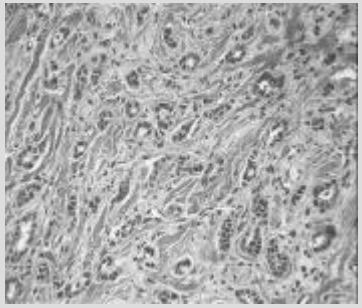
Wir messen die momentane Neuproduktion von Proteinen und nicht den Vorrat!

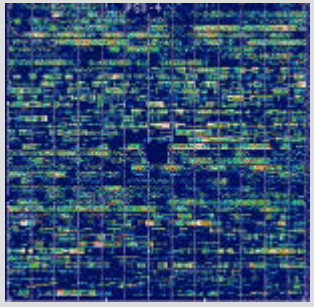
Wir protokollieren was die Zelle gerade tut!



Der Microarray ist eine Momentaufnahme des Arbeitsprozesses in den Zellen

DNA Chip





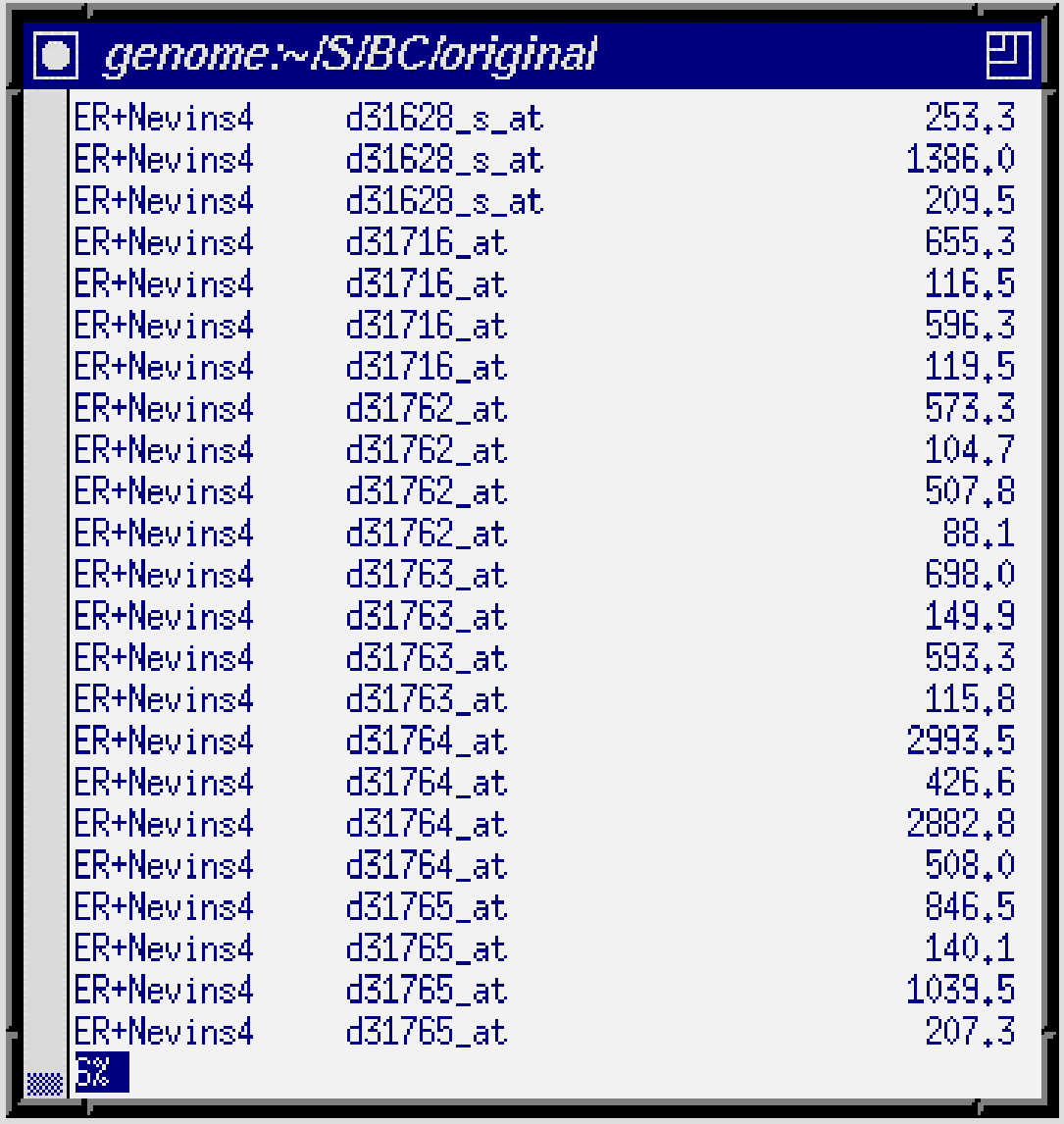
Gewebe

genome:~/SIBCloriginal

ER+Nevins4	d31628_s_at	253.3
ER+Nevins4	d31628_s_at	1386.0
ER+Nevins4	d31628_s_at	209.5
ER+Nevins4	d31716_at	655.3
ER+Nevins4	d31716_at	116.5
ER+Nevins4	d31716_at	596.3
ER+Nevins4	d31716_at	119.5
ER+Nevins4	d31762_at	573.3
ER+Nevins4	d31762_at	104.7
ER+Nevins4	d31762_at	507.8
ER+Nevins4	d31762_at	88.1
ER+Nevins4	d31763_at	698.0
ER+Nevins4	d31763_at	149.9
ER+Nevins4	d31763_at	593.3
ER+Nevins4	d31763_at	115.8
ER+Nevins4	d31764_at	2993.5
ER+Nevins4	d31764_at	426.6
ER+Nevins4	d31764_at	2882.8
ER+Nevins4	d31764_at	508.0
ER+Nevins4	d31765_at	846.5
ER+Nevins4	d31765_at	140.1
ER+Nevins4	d31765_at	1039.5
ER+Nevins4	d31765_at	207.3

3%

Expressions-Profil



Ein Expressions-Profil ist eine lange Liste von Zahlen

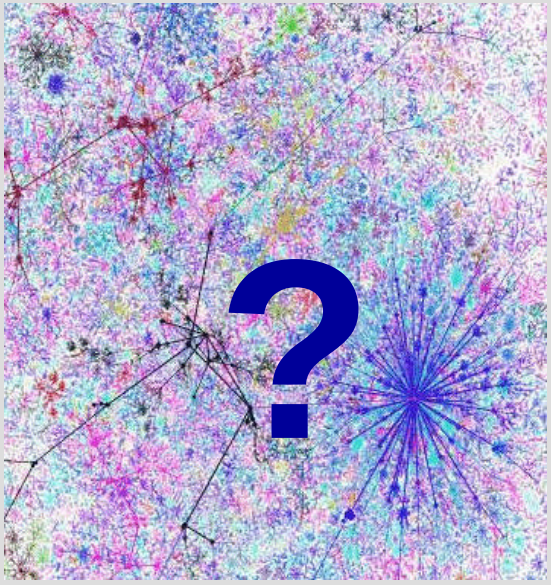
Für jedes Transcript eine Expressions-Intensität

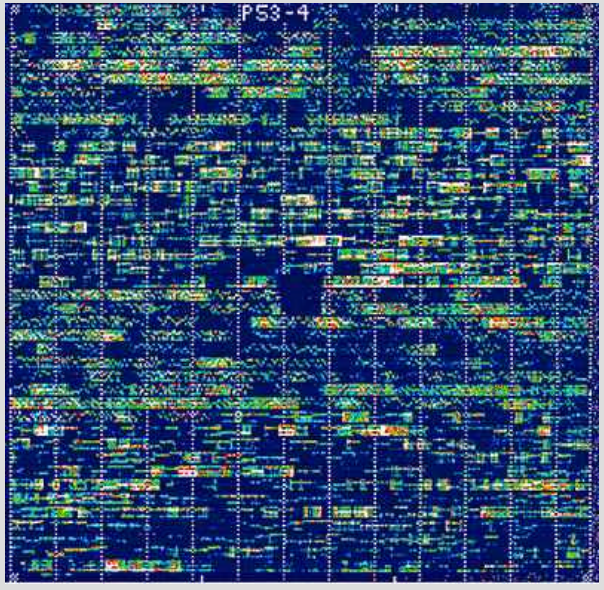
Das Profil gewährt einen Blick in die Zellen der Gewebeprobe

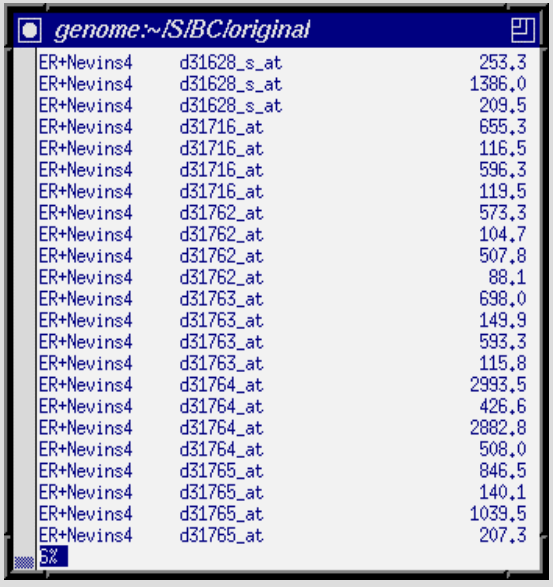
Es ist ein sehr komplexer diagnostischer Befund

Tagträumerei -oder-

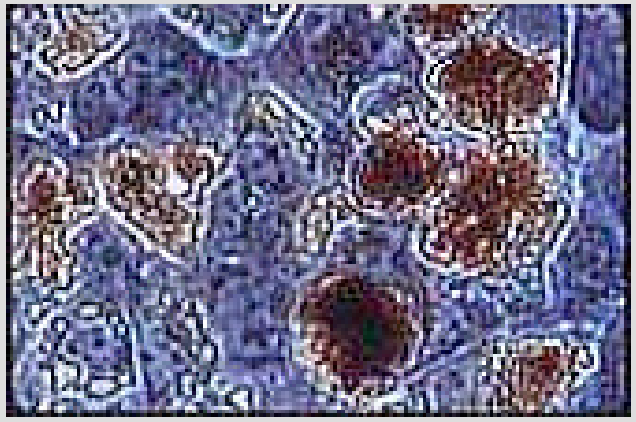
Wohin soll das führen?







z.B. Krebs

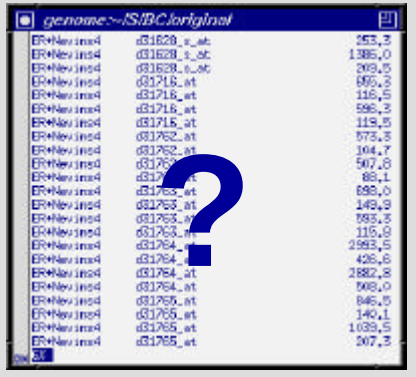


Typische Kriterien die bei der Diagnose und Therapie eine Rolle spielen:

Größe, Lage, Zellmorphologie, Ursprungsgewebe, Differentiationsstatus, Mutationen, etc.

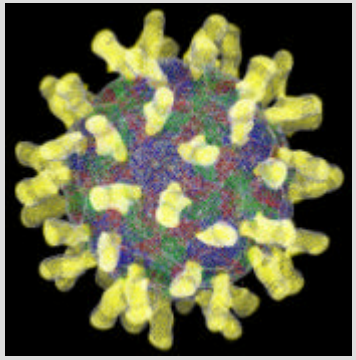
Interessant wäre (außerdem) :

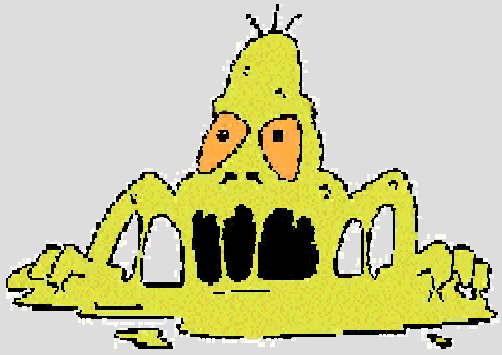
Arbeiten einige Zellen bereits an der Ausbildung von Metastasen?

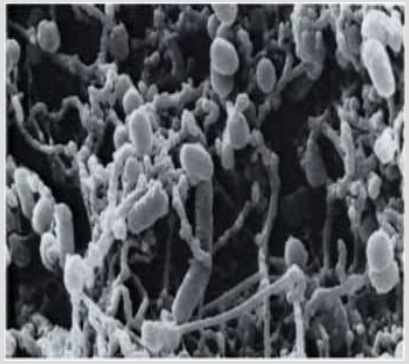


z.B. Infektionen

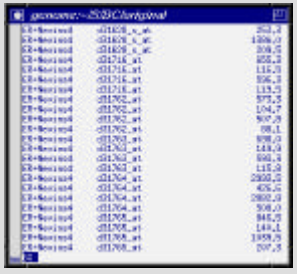
Systematik basierend auf Eigenschaften der Erreger

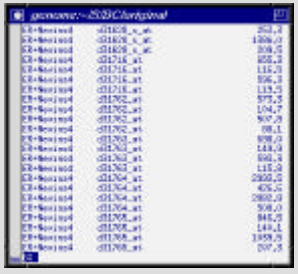


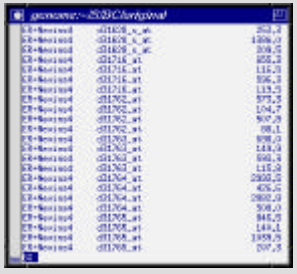




Systematik basierend auf der Zell- (Immun-) Antwort







Es ist also interessant Expressionsprofile von Patienten zu erheben und zu analysieren ... und das wird auch getan ... zum Beispiel im Rahmen des ...

Nationales
Genomforschungsnetz

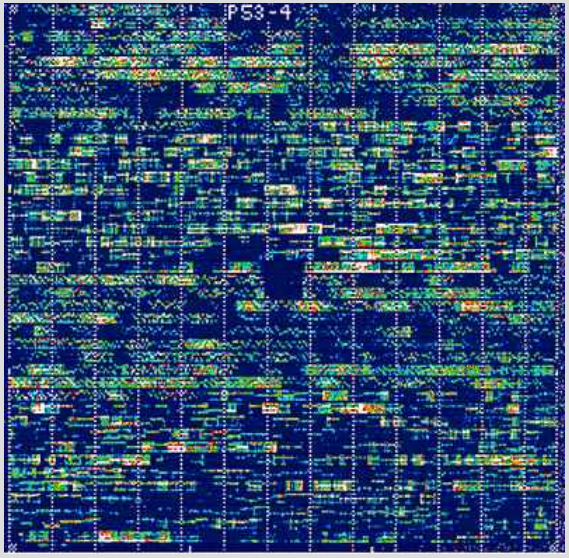
Wir sind eine Bioinformatikgruppe am

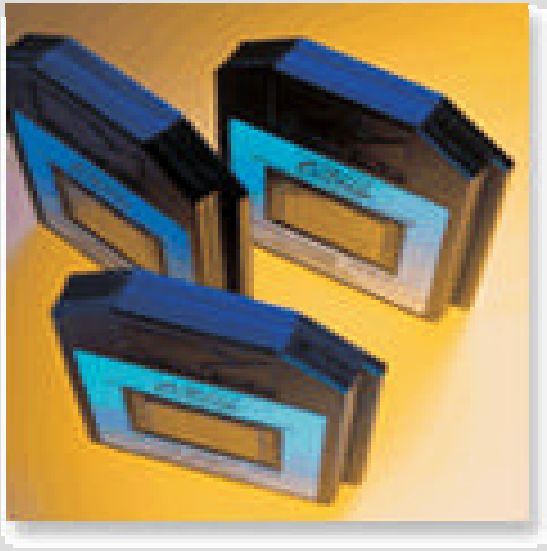
MPIMG, die an diesen Projekten mitarbeitet.

Wozu braucht man dabei Bioinformatik?

Probleme gibt es genug ...

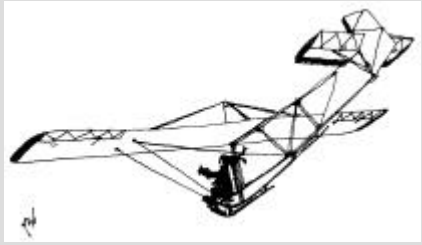
z.B. in der Technologieentwicklung

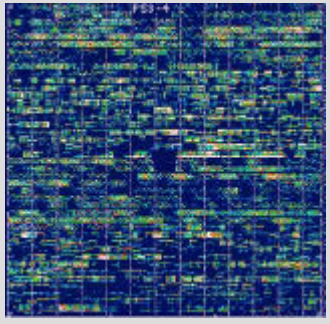




- Chip Design
- Bildverarbeitung
- Qualitätskontrolle
- Skalierung
- Normalisierung, ...

Dies sind Probleme die sich mit der Weiterentwicklung der Technologie ändern und teilweise erübrigen werden ... trotzdem sind sie heute von entscheidender Bedeutung!





Es ist eine neue Technologie:

-schlechte Datenqualität,Rauschen

-Artefakte

-kaputte Microarrays

-wenig Erfahrung im Umgang mit der Technologie

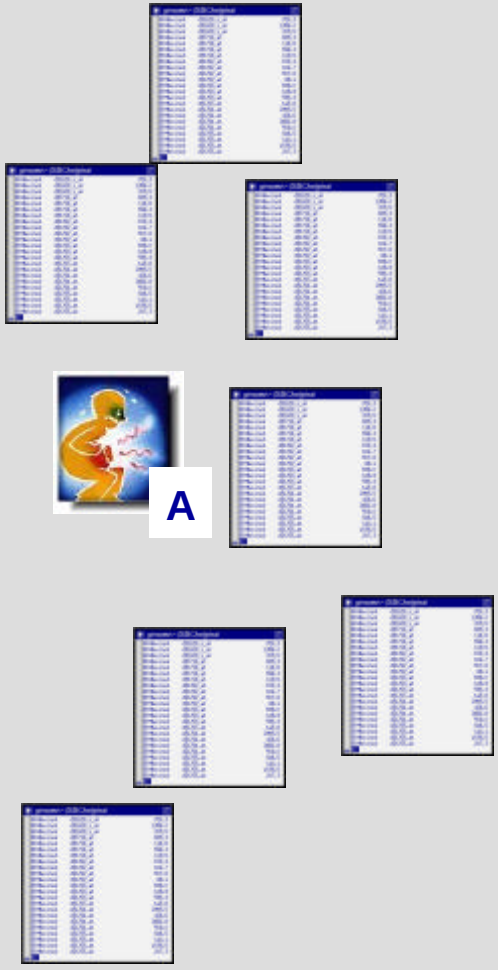
?

Es gibt aber auch ein Problem, daß sich auch bei perfekter Technologie immer noch auftreten wird ...

... ein zeitloses Kernproblem das mit der Idee Diagnosen auf sehr komplexe Befunde aufzubauen immanent verbunden ist ...

... und darüberhinaus ein generelles Problem beim Analysieren hochdimensionaler Daten ...

Ausgangs-situation



Zwei Entitäten

A und B

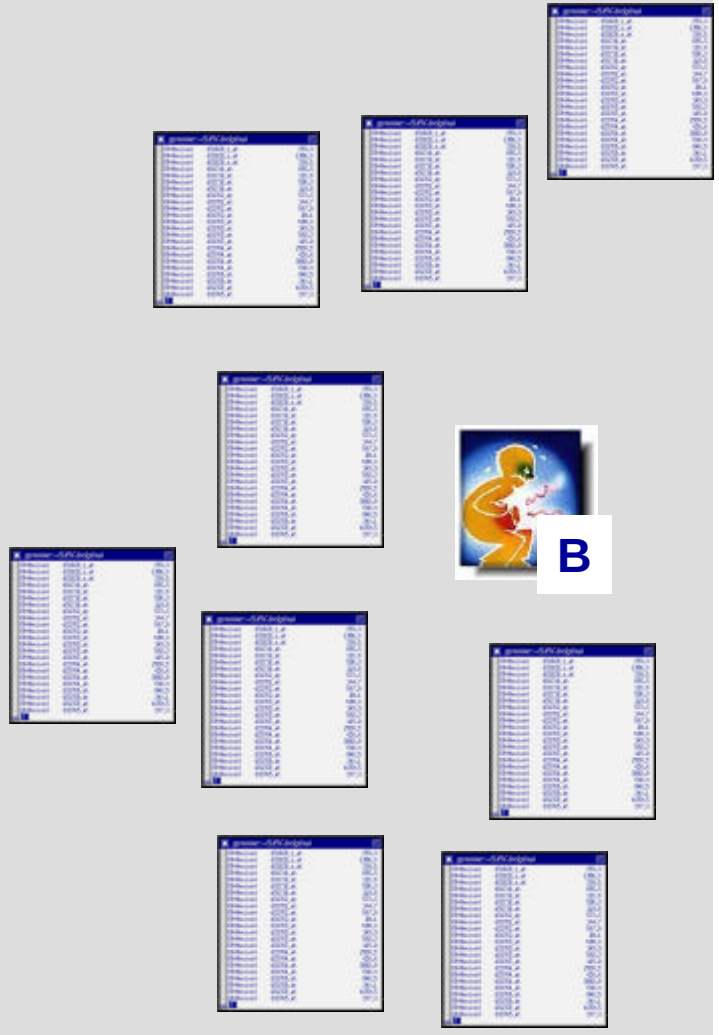
z.B.

-gutartiger Tumor vs. bösartiger Tumor

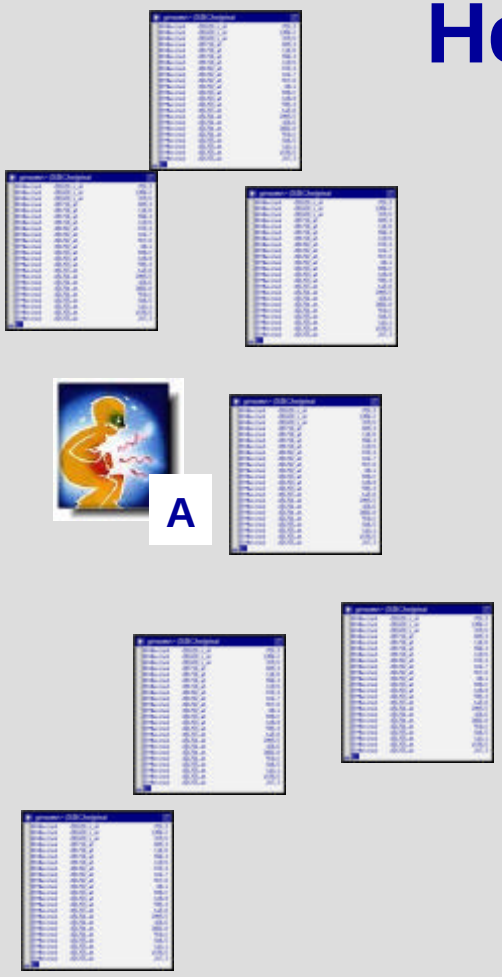
-Medikament erfolgreich vs. Medikament nicht erfolgreich

-etc.

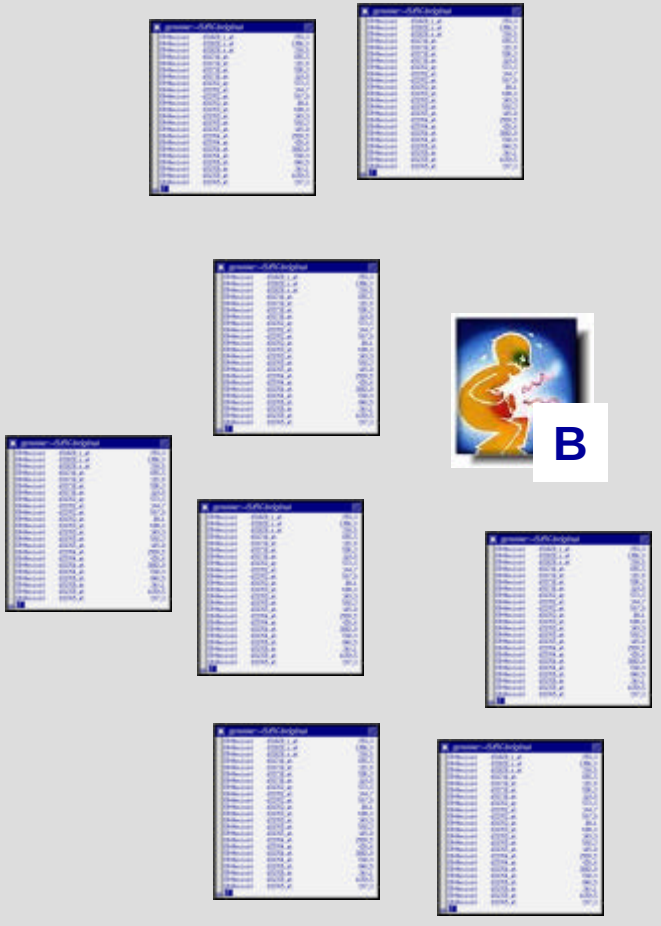
Expressionsprofile von Patienten beider Entitäten



Naheliegende (aber naive) Herangehensweise



Suche nach Unterschiede in der Genexpression von Typ A Patienten zu Typ B Patienten



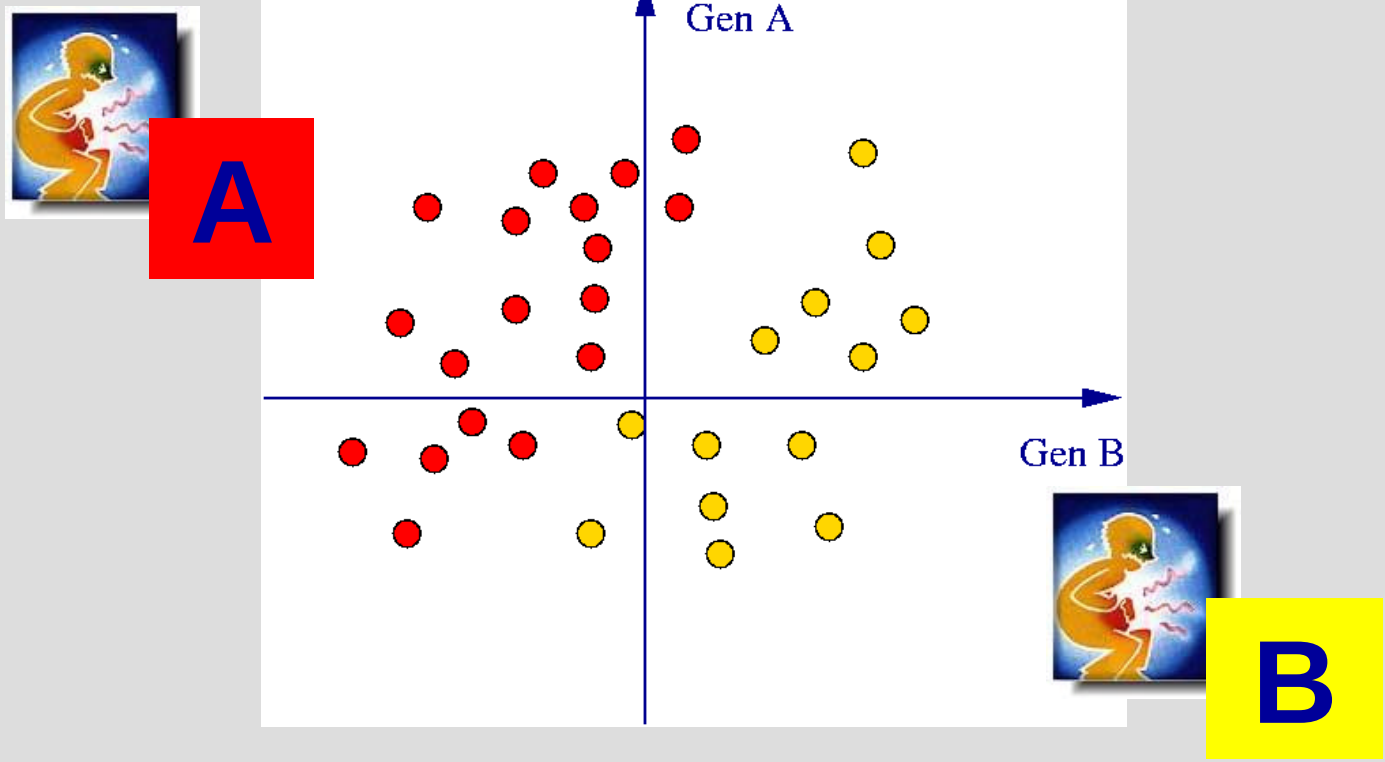
Was ist daran naiv?

10.000 Gene sind etwas unübersichtlich

Also beschränken wir uns für den Augenblick auf zwei

Gen A und Gen B

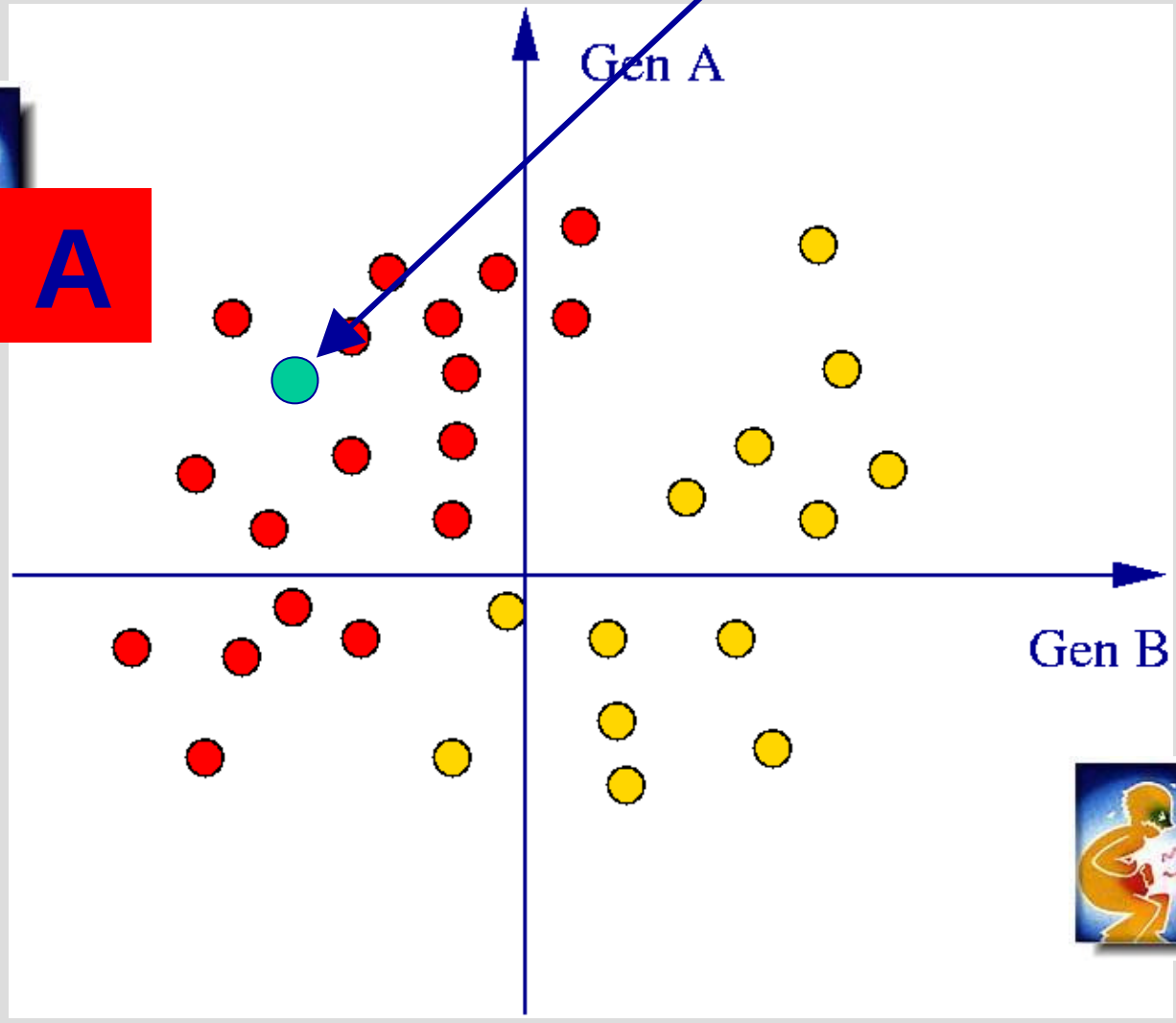
Angenommen wir hätten nur die Expressionswerte von zwei Genen



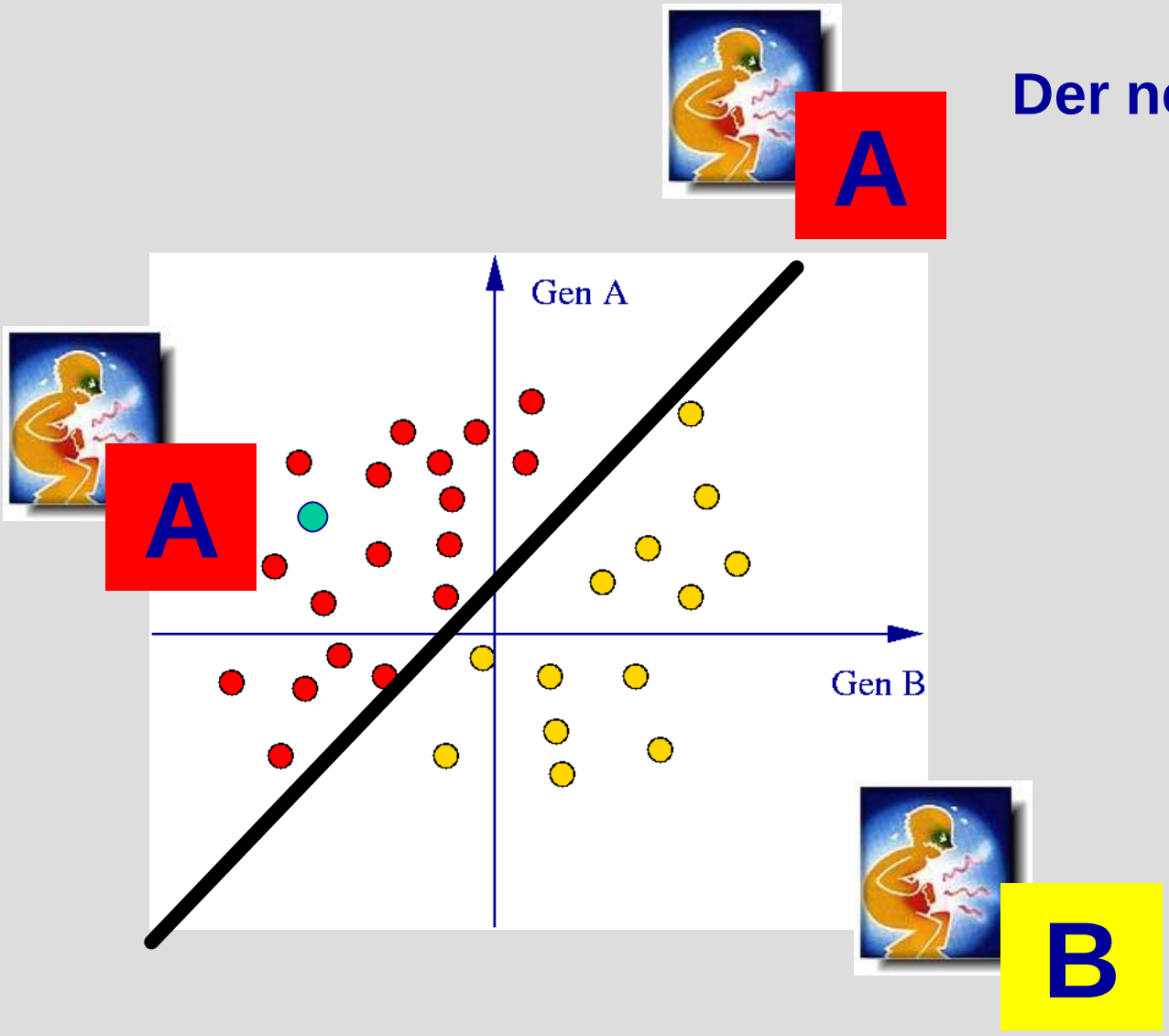
Ja, es gibt einen Unterschied

Ein neuer Patient

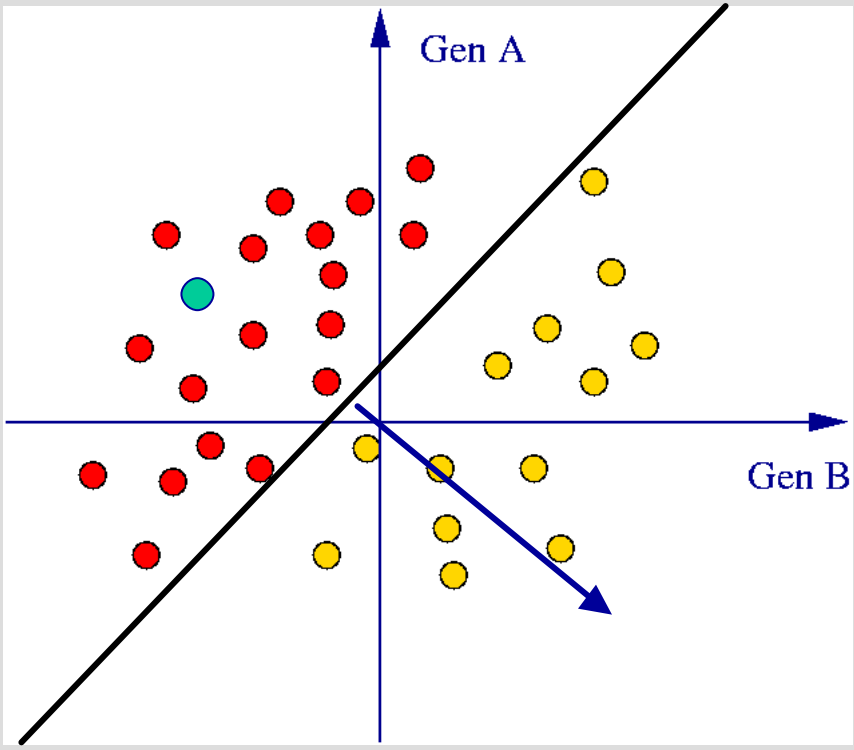
A



B



In dieser Situation ist alles klar.



Berechne den Normalenvektor einer trennenden Gerade, dieser ist dann die diagnostische Signatur

... die trennende Gerade ist nicht eindeutig

Allgemeiner:

Was genau soll eine Signatur sein?

$x_1, \ldots, x_{30000}$: Expressionslevel

$f(x_1, \ldots, x_{30000})$: Funktion die den Expressionslevel eine Zahl zuordnet

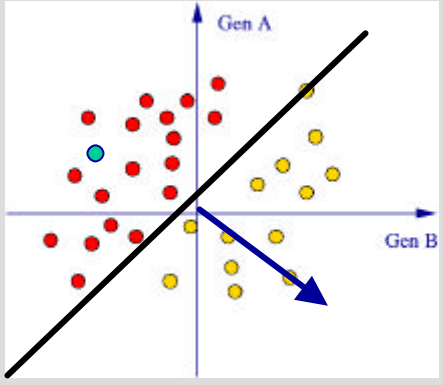
Hohe Werte von f sprechen für Klasse 1
niedrige für Klasse 2

Zum Beispiel:

$$f(x_1, \ldots, x_{30000}) = x_1$$ **Gen 1 ist die Signatur**

Oder, ein Normalenvektor ist die Signatur:

$$f(x_1, \ldots, x_{30000}) = b_0 + b_1 x_1 + b_2 x_2$$

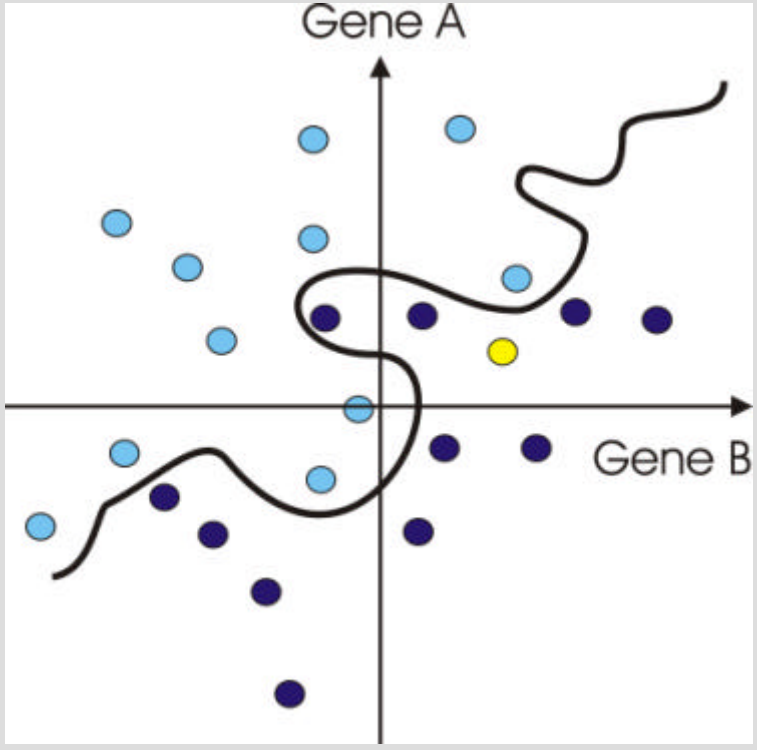


Falls x1 und x2 die beiden Gene im Diagramm sind

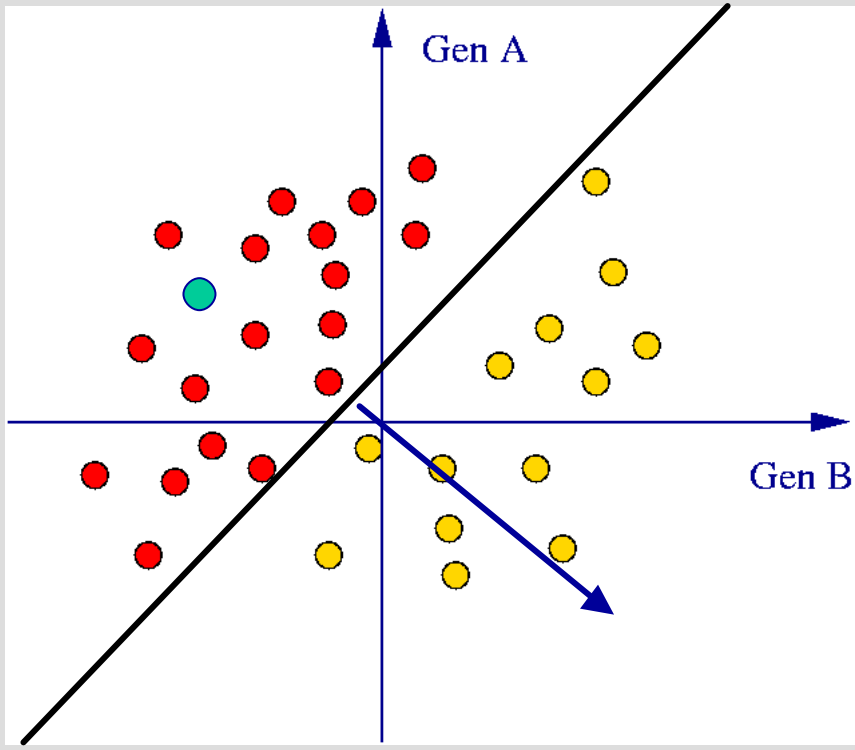
Das gleiche mit allen Genen ergibt dann:

$$f(x_1, \ldots, x_{30000}) = b_0 + \sum_{i=1}^{30000} b_i x_i$$

Oder man nimmt eine sehr komplizierte Signatur



$$f(x_1, \ldots, x_{30000}) = \text{kompliziert}$$

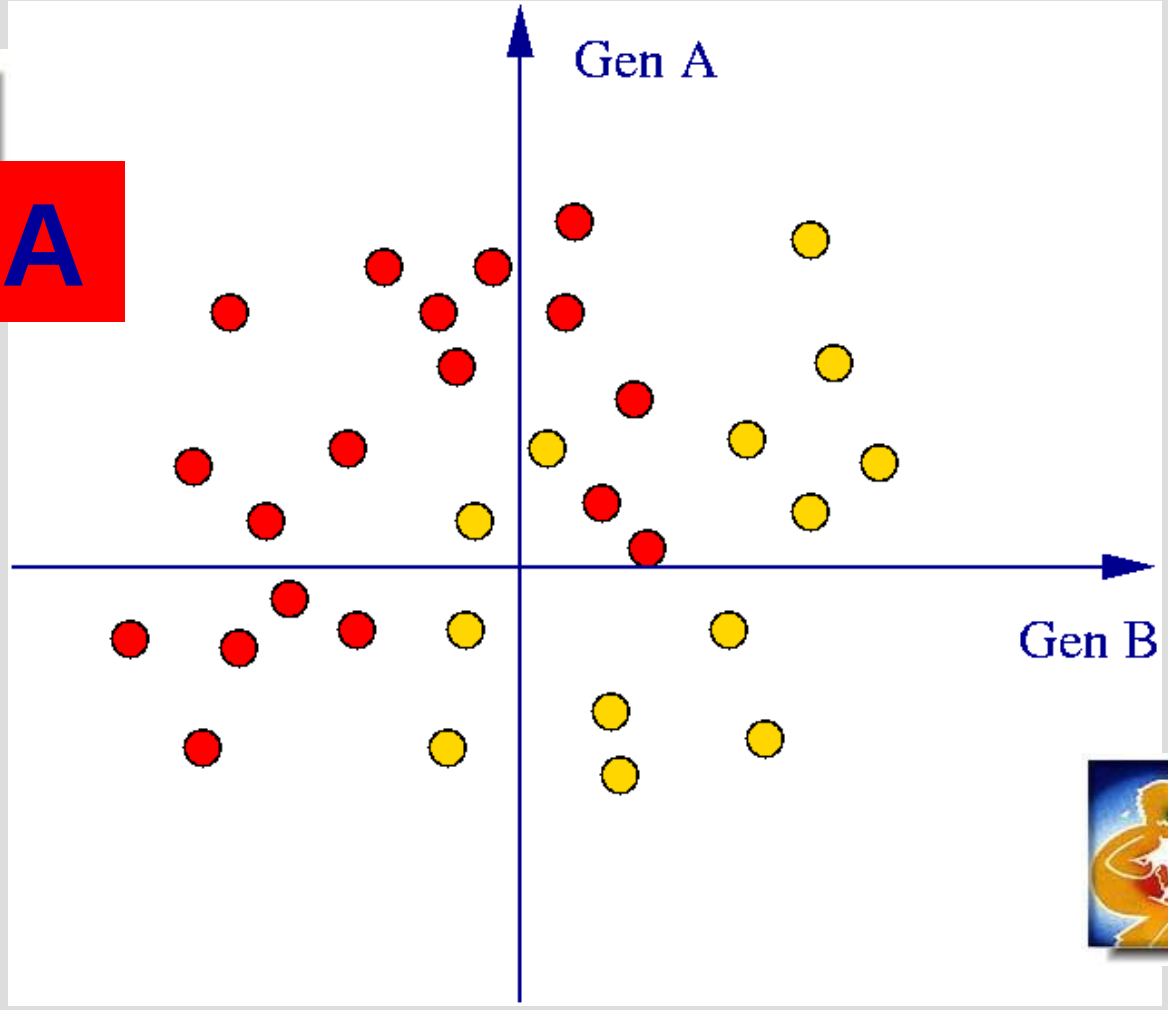


Leider sehen Expressionsdaten nie so schön aus

Was kann schief gehen?

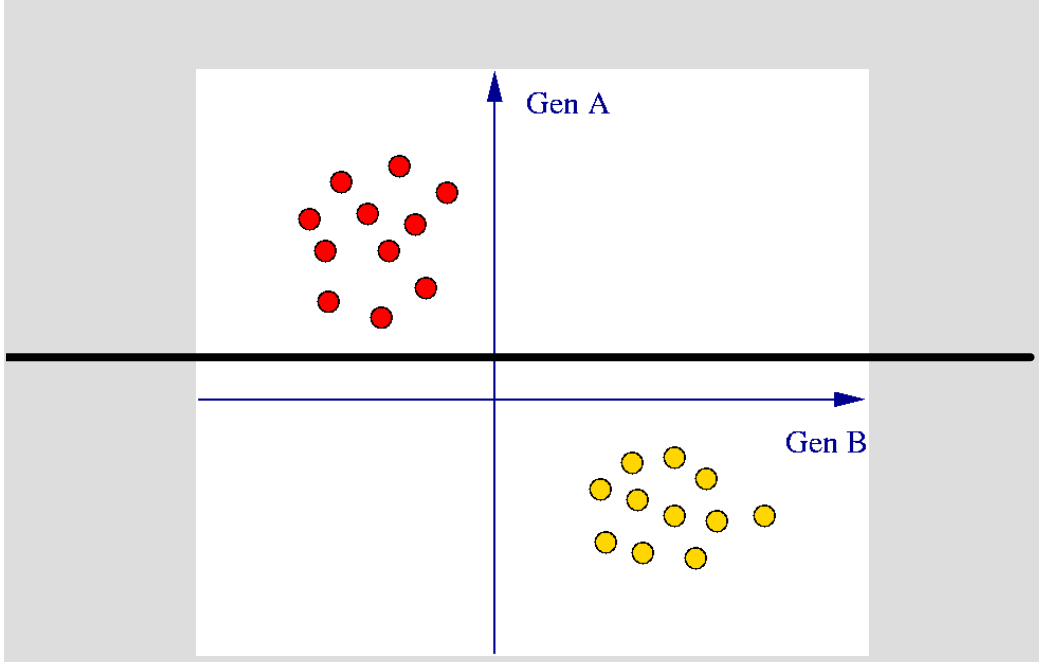
Es gibt keine Gerade, die die Gruppen trennt

A



B

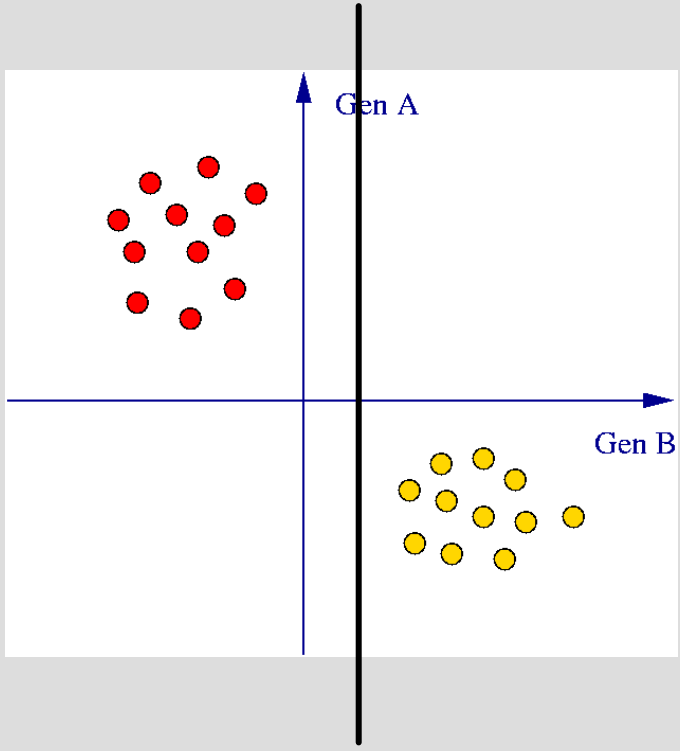
Gen A ist wichtig



Gen A hoch

Gen A niedrig

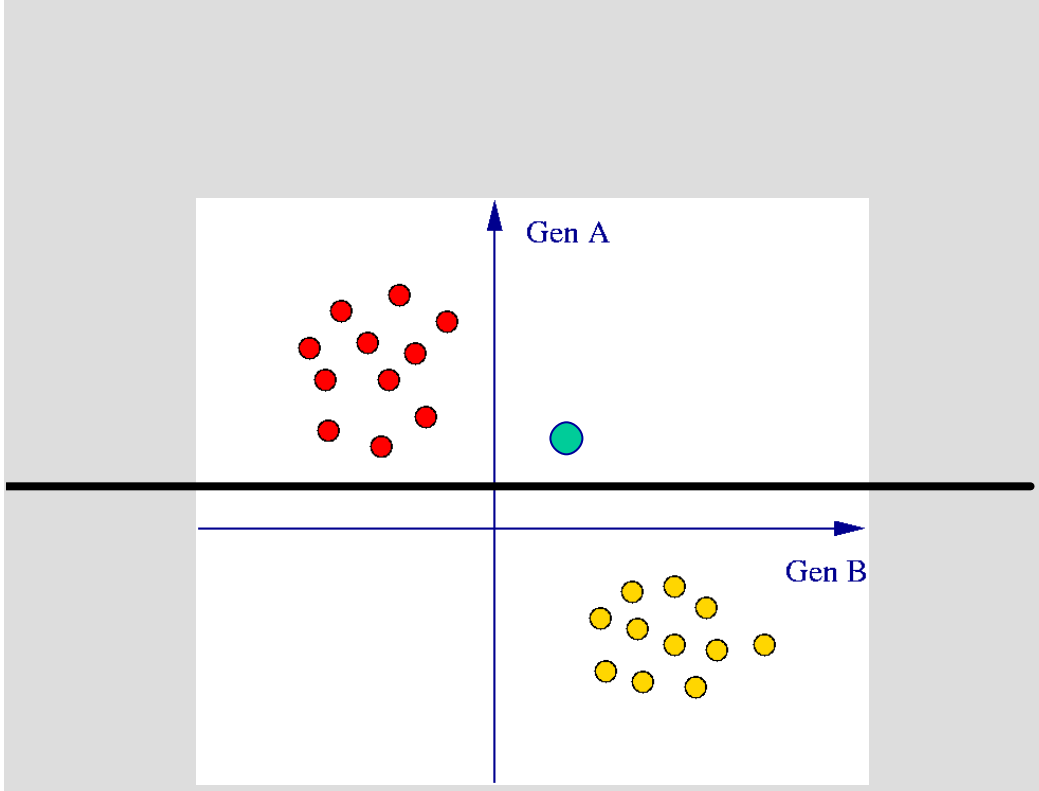
Gen B ist wichtig



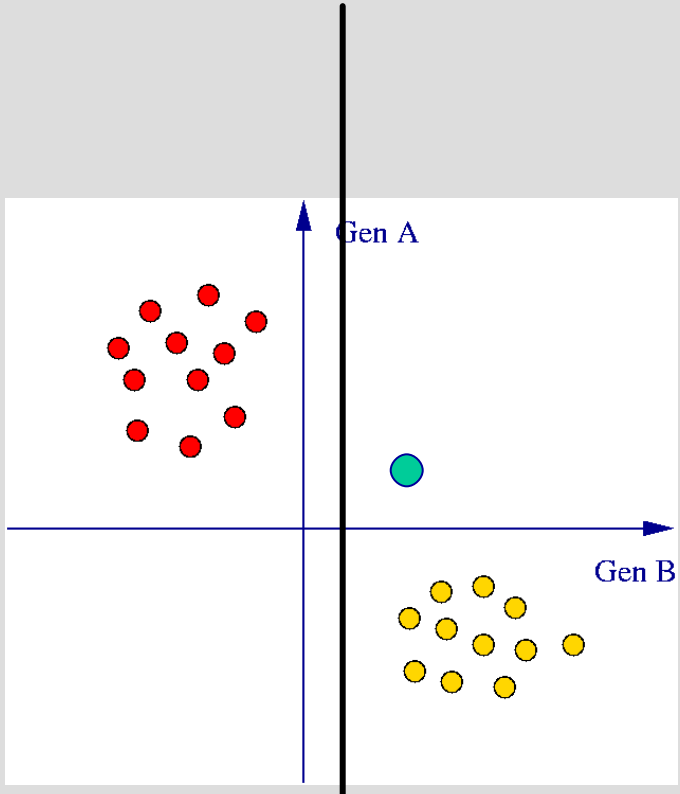
Gen B niedrig

Gen B hoch

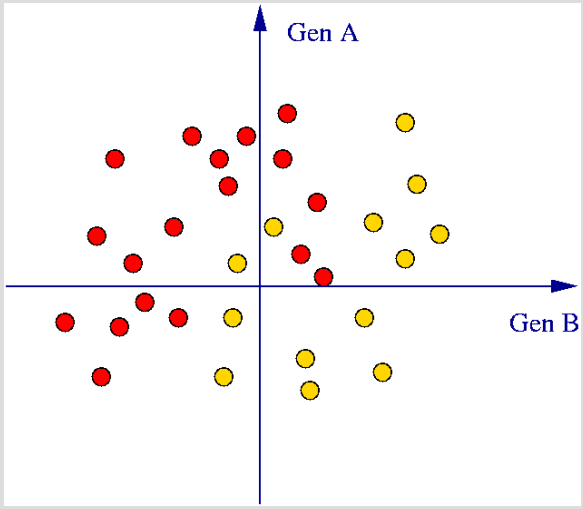
Neuer Patient ?



A

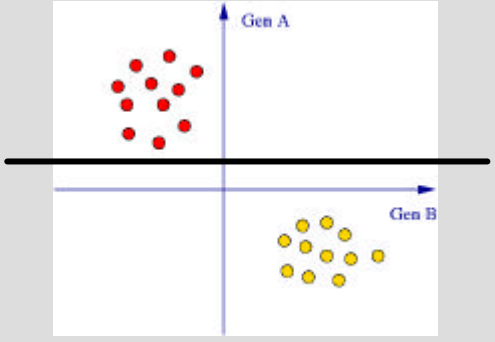


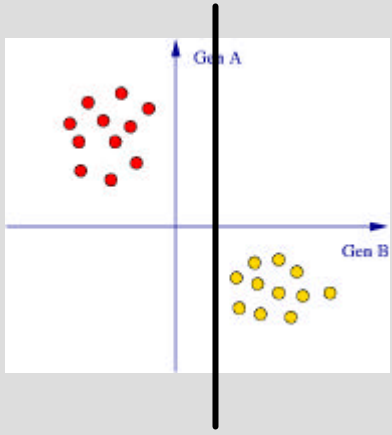
B



Problem 1:

Keine Gerade

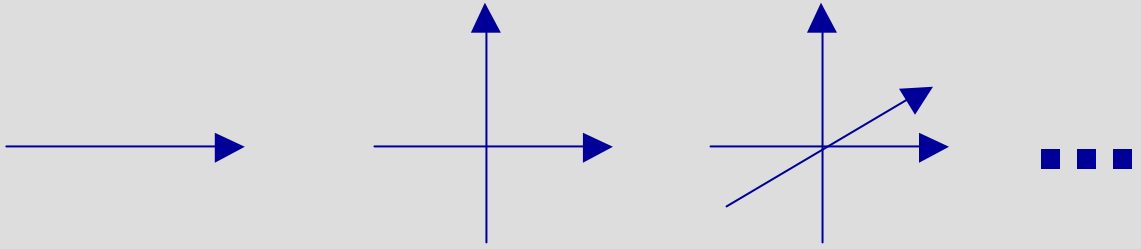




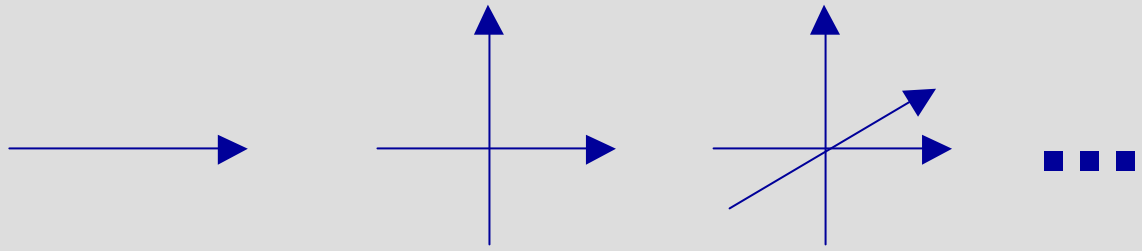
Problem 2:

Zuviele unterschiedliche Geraden

In der Praxis untersuchen wir tausende Gene und im allgemeinen mehr Gene als Patienten



Und in den Weiten eines 30000 dimensionalen Raumes herrschen andere Gesetze



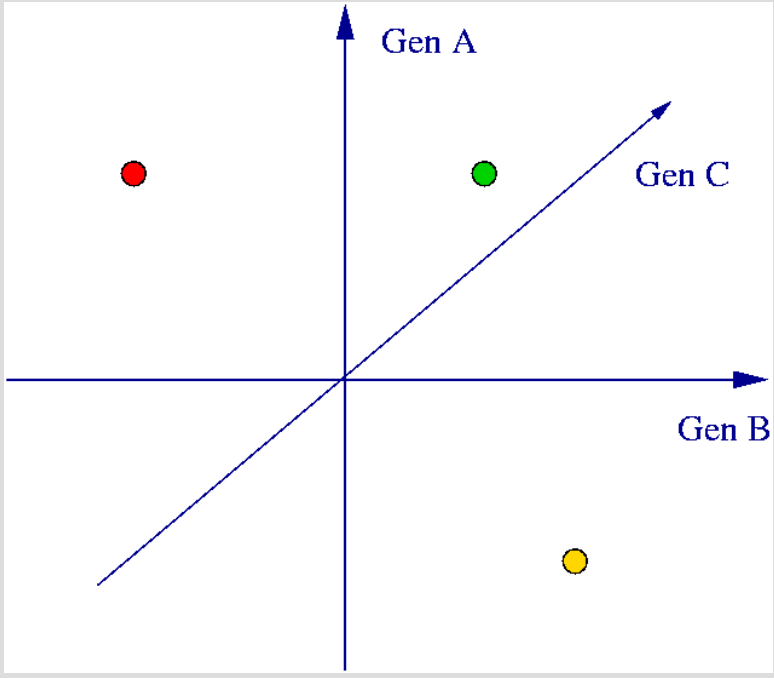
1 2 3

30000

- **Problem 1 entsteht nie!**
- **Problem 2 entsteht praktisch immer!**

Überlegen Sie sich das einmal kurz in drei Dimensionen

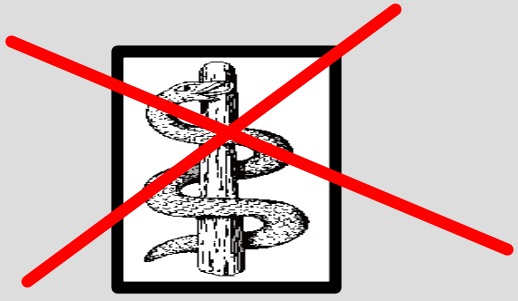
Also für drei Gene, zwei Patienten mit bekannter Diagnose, einem neuen Patienten und einer Trennebene statt einer Trenngerade

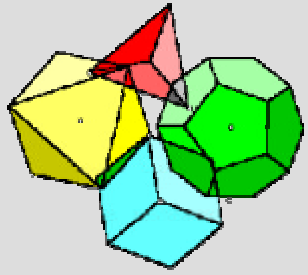


OK! Wenn alle Punkte auf einer Geraden liegen, geht es nicht immer. Das ist bei Messungen aber sehr unwahrscheinlich und kommt praktisch nie vor.

Aus den Daten alleine kann man weder feststellen, welche Gene wichtig für die Diagnose sind, noch kann man zweifelsfrei eine Diagnose für den nächsten Patienten stellen.

Dieses Problem hat mit Medizin wenig zu tun. Es ist ein geometrisches Problem.





Es gibt also auf alle Fälle lineare Signaturen die:

- gut artige von bösartigen Tumoren unterscheiden

- oder bei beliebiger Anordnung der Patienten die mit gerader Nummer von denen mit ungerader Nummer

Im wesentlichen heißt das Auffinden von Unterschieden in der Genexpression zweier Patientengruppen nichts!

Langsam!

Es gibt aber durchaus auch bedeutungsvolle Unterschiede in der Genexpression dieser Krankheitentitäten und die muß man auf dem Chip auch sehen können

Es gibt also auf der einen Seite völlig bedeutungslose Signaturen, auf der anderen Seite aber auch welche die tatsächliche Disregulation widerspiegeln.

Wie kann man die beiden Fälle unterscheiden?

Woran kann man die bedeutungslosen Signaturen erkennen?

Sie treten in großen Mengen auf

Ihre Parameter haben eine hohe Varianz

Unterbestimmtheit

Sie spiegeln Details wieder aber nicht das Wesentliche

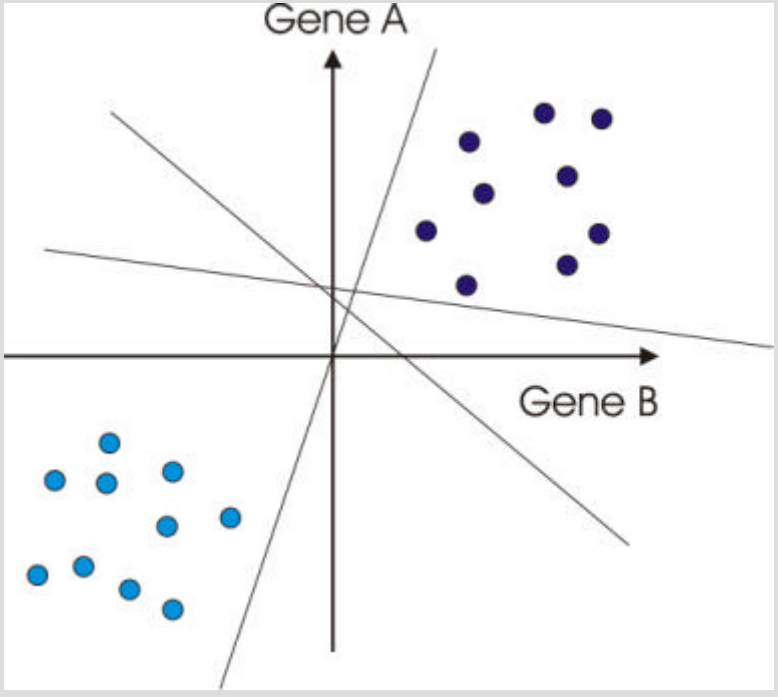
Sie lernen auswendig aber abstrahieren nicht

Overfitting

Unterbestimmtheit

Sie treten in großen Mengen auf

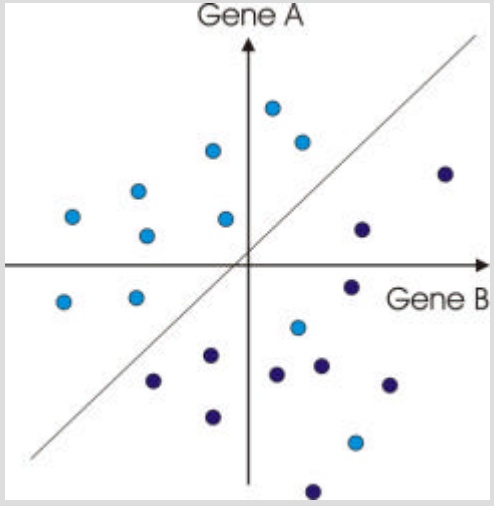
Ihre Parameter haben eine hohe Varianz



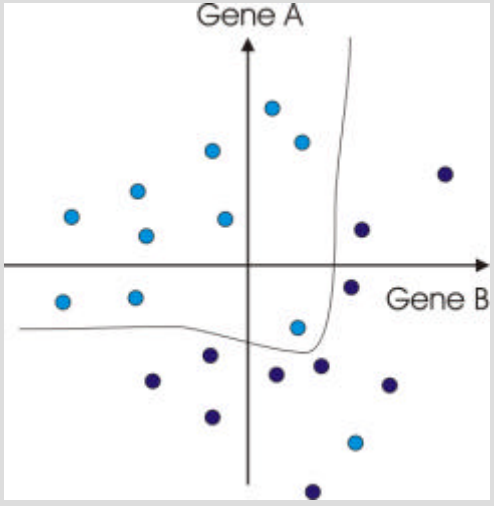
Overfitting

Sie spiegeln Details wieder aber nicht das Wesentliche

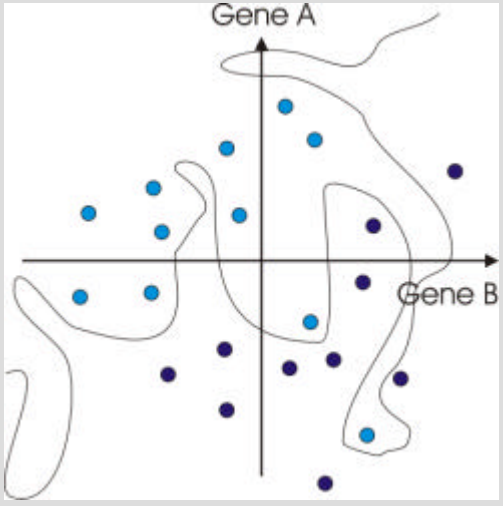
Sie lernen auswendig aber abstrahieren nicht



2 Fehler



1 Fehler



keine Fehler

Signaturen müssen nicht perfekt sein

Welche Strategien gibt es gute Signaturen zu bekommen ?

Z.B. ...

- **Genselektion gefolgt von linearer Klassifikation**
- **Support Vector Machines**

Worauf beruhen diese Verfahren?

Genselektion

sieht man zurzeit am häufigsten

Wenn wir alle Trennebenen betrachten gibt es immer eine die eine perfekte Signatur ist, ohne daß es einen biologischen Grund dafür geben muß

Betrachten wir aber nur Ebenen deren Lagen von maximal 20 Genen abhängen, so gibt es darunter nicht unbedingt immer eine perfekte Signatur, gibt es sie doch, sind die Chancen gut, daß es dafür einen biologischen Grund gibt

Wählen wir die Gene so aus, daß jedes für sich ein „guter“ Marker ist, schränkt das die Menge möglicher Signaturen weiter ein

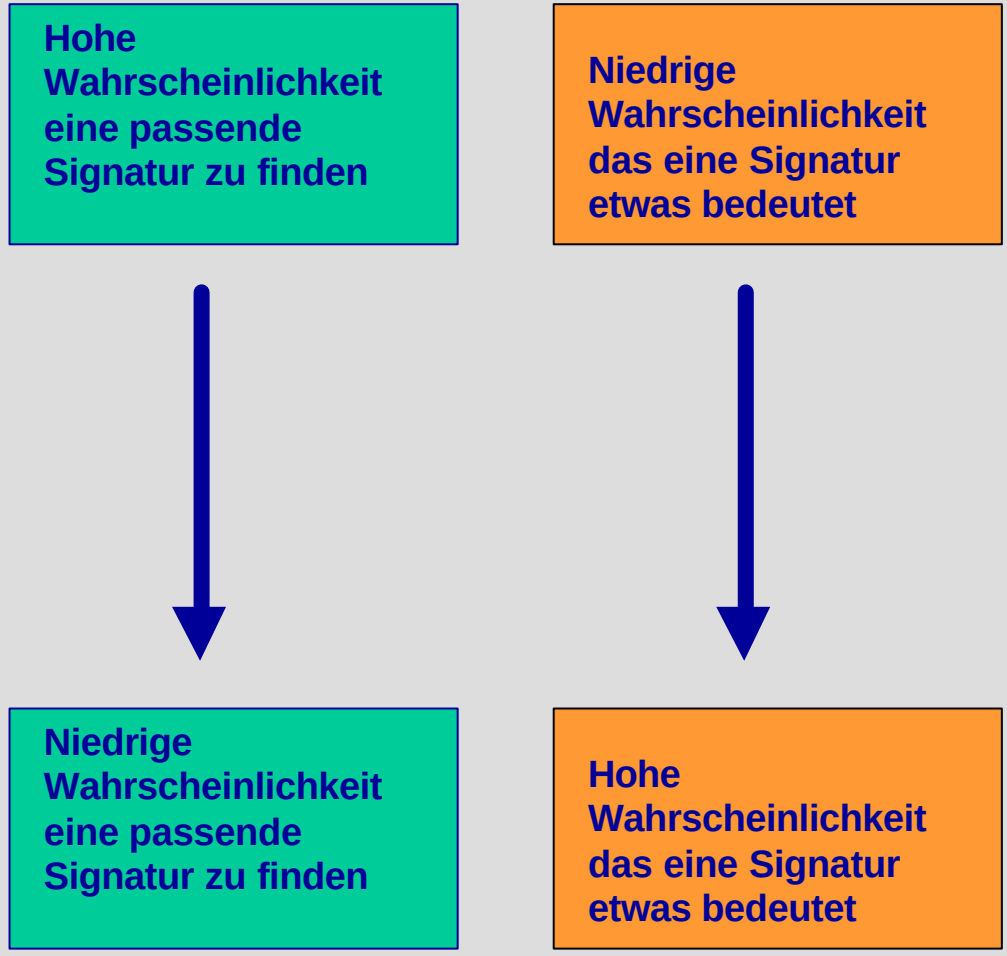
Beispiele für Mengen möglicher Signaturen

- Alle quadratischen Trennflächen

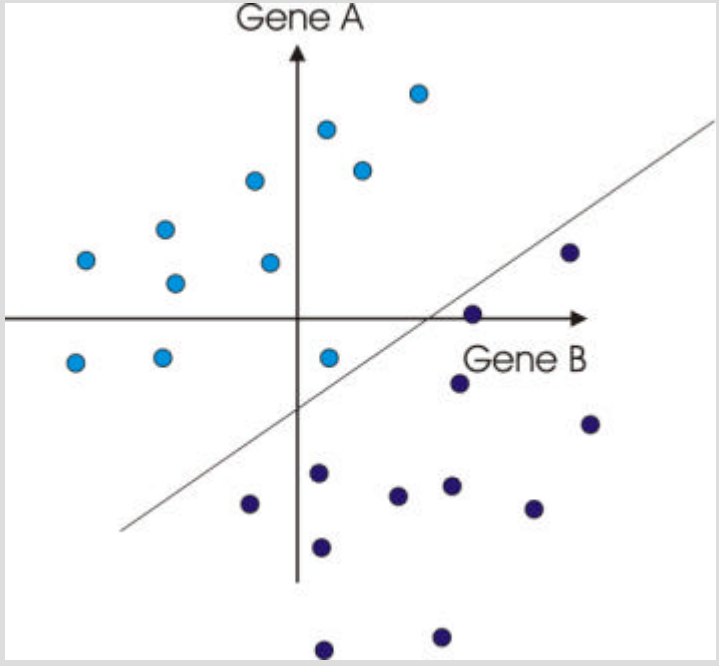
- Alle Trennebenen

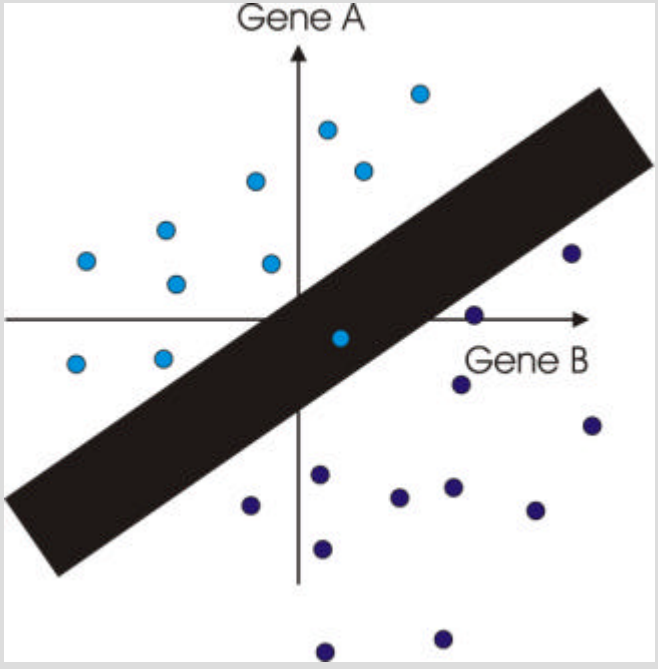
- Alle Trennebenen die von 20 Genen oder weniger abhängen

- Alle Trennebenen die von 20 vorgegebenen Genen abhängen



Support Vector Machines





Dicke Trennebenen: Mit einer dünnen Trennebene lassen sich die Daten immer trennen. Aber nicht unbedingt mit einer dicken. Large Margin Classifiers

Sowohl Genselektion als auch Support Vector Machines engen die Menge möglicher Signaturen a priori ein, wenn auch auf unterschiedliche Art und weise.

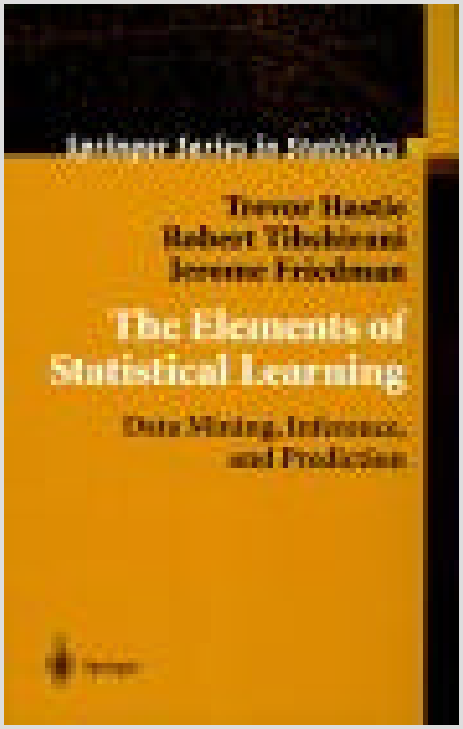
Geneselektion will wenig Gene in der Signatur

SVM wollen einen gebührenden Abstand der Daten zur trennenden Ebene

Es gibt noch viele andere Strategien ....

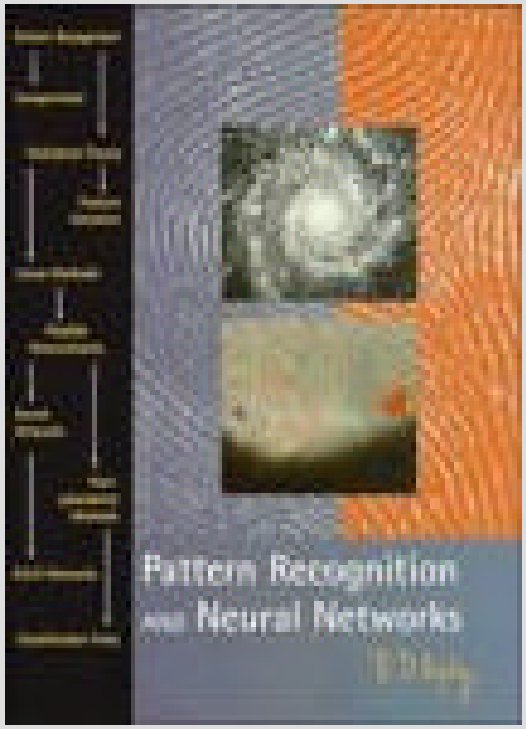
Lernverfahren

Ridge Regression, LASSO, Kern-Basierte-Methoden, Additive Modelle, Klassifikationsbäume, Bagging, Boosting, Neuronale Netze, Support Vector Machines, Relevance Vector Machines, Nearest-Neighbors, Transduction etc. etc.



The Elements of Statistical Learning

Hastie, T. Tibshirani, R. Friedman, J



Pattern Recognition and Neural Networks

Brian D. Ripley