

**Einführung in die Angewandte Bioinformatik:  
Analyse und Design von DNA Microarrays  
09.07.2009**

Prof. Dr. Sven Rahmann

## Transcript – Omics (Genexpressionsanalyse)

Zum Verständnis von lebenden Organismen untersucht man u.a. ihr

- Genom
- Transkriptom
- Proteom
- Metabolom,

sowie Wechselwirkungen zwischen diesen Ebenen.

Das Genom wird durch DNA-Sequenzierung in Genomprojekten bestimmt.  
Das Proteom und Metabolom z.B. mit Massenspektrometrie.

## Transcript – Omics (Genexpressionsanalyse)

Das Transkriptom (Gesamtheit der in RNA transkribierten DNA) wird bestimmt mittels

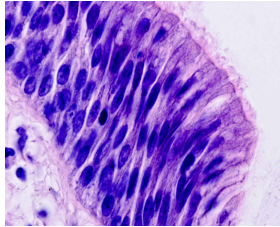
- aktuell: DNA-Microarrays,
- Zukunft: Hochdurchsatz-Sequenzierung von mRNA.

Transkription variiert

- von Gewebe zu Gewebe,
- abhängig von äußeren und inneren Einflüssen,
- als Antwort auf Signale anderer Zellen, ...

## Genexpressionsanalyse: Motivation Beispiel Bronchialkarzinom

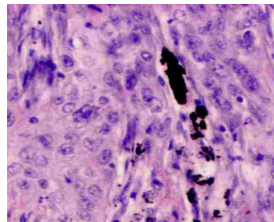
Phänotyp „gesund“



Proteinexpression



Genexpression



Differenzielle  
Proteinexpression



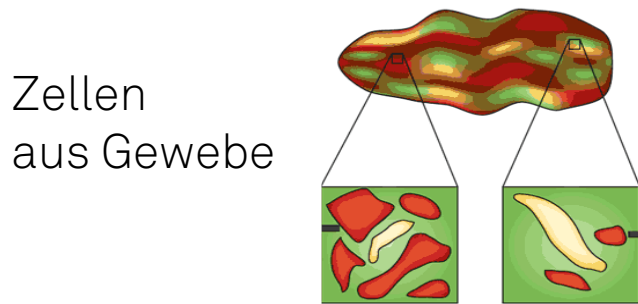
Differenzielle  
Genexpression

Phänotyp „Bronchialkarzinom“

**Kandidaten-Gen:** differenziell exprimiert in Tumorzellen.  
Identifikation wichtig für Diagnostik und Therapie.

## (Alte) cDNA-Array-Technologie

### 2. Vorbereitung der Proben (samples)



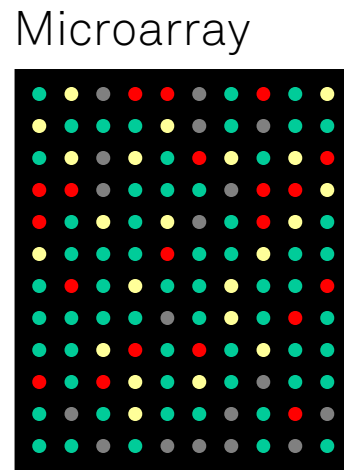
Zellen  
aus Gewebe

mRNA-Extraktion  
& Amplifikation

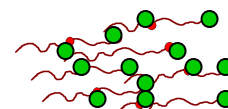
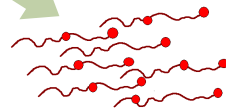


Markierung mit  
Fluoreszenz-Farbstoffen

### 3. Hybridisierung



Microarray



### 1. Vorbereitung des Microarrays

PCR-Amplifikation der  
zu untersuchenden Gene  
(pro Gen ein Arbeitsschritt)



Spotten der cDNA  
auf definierten Plätzen



### 4. Bild- und Datenanalyse

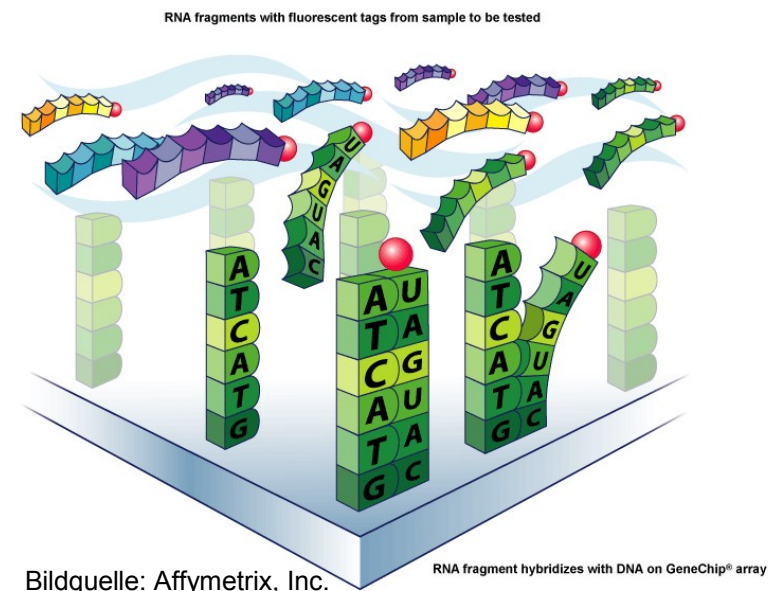
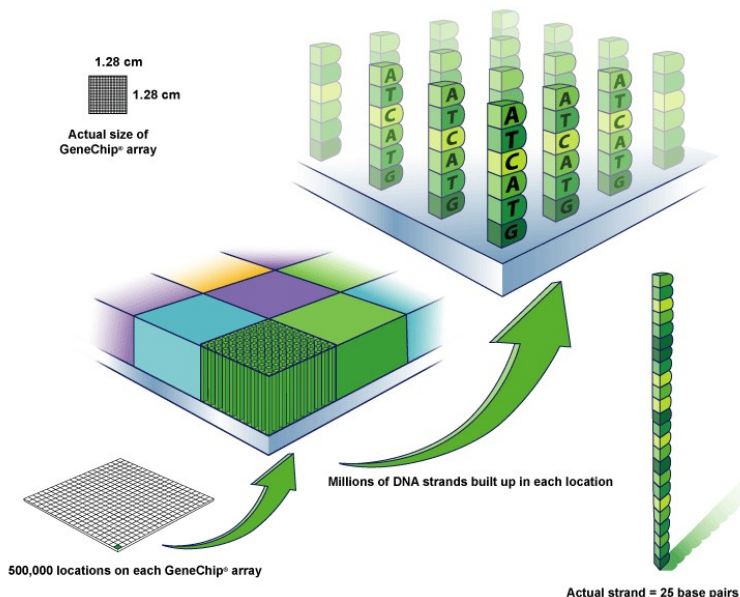
Rot/Grün-Verhältnis  
Hintergrundkorrektur

## (Aktuelle) Oligonukleotid-Array-Technologie

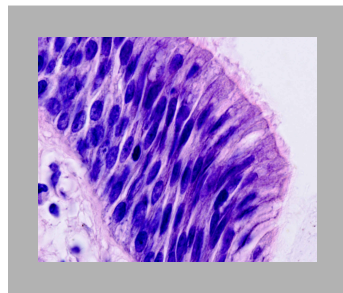
Oligonukleotid-Microarrays bestehen aus Mio. **spots** oder **features**.

- Spot enthält Mio. Kopien eines bestimmten Oligonukleotids.
- Oligonukleotid hybridisiert (spezifisch) an mRNA / cDNA (Watson-Crick-Paarung).
- mRNA/cDNA ist fluoreszierend markiert.
- Ungebundene RNA wird abgewaschen.
- Fluoreszenz-Signal auf Spot erlaubt Rückschluss auf Genexpression.

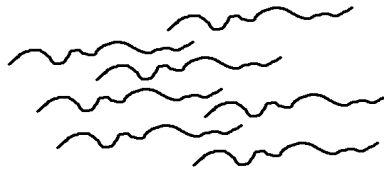
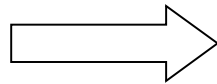
Wichtig: Oligonukleotid-Sonden **gen-spezifisch**.



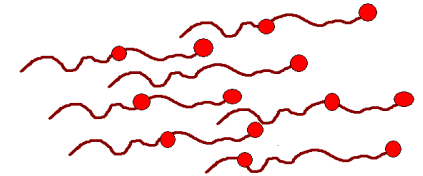
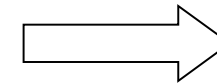
## Ablauf eines einzelnen Microarray-Experiments



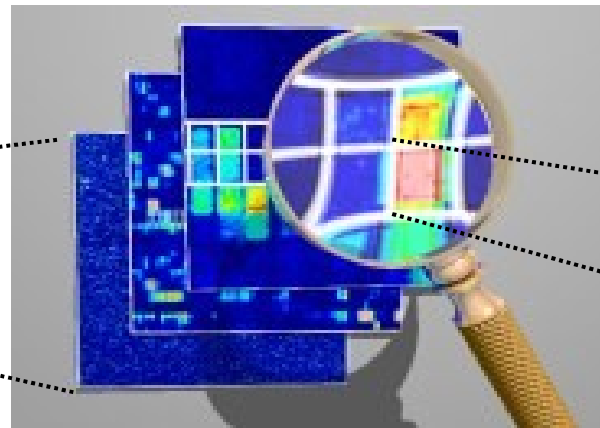
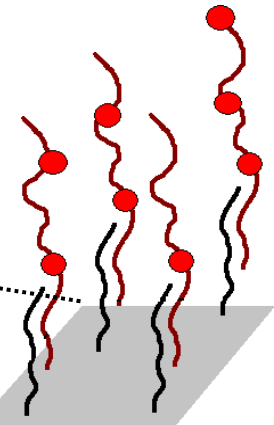
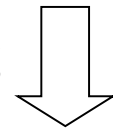
Extraktion der poly-A-RNA



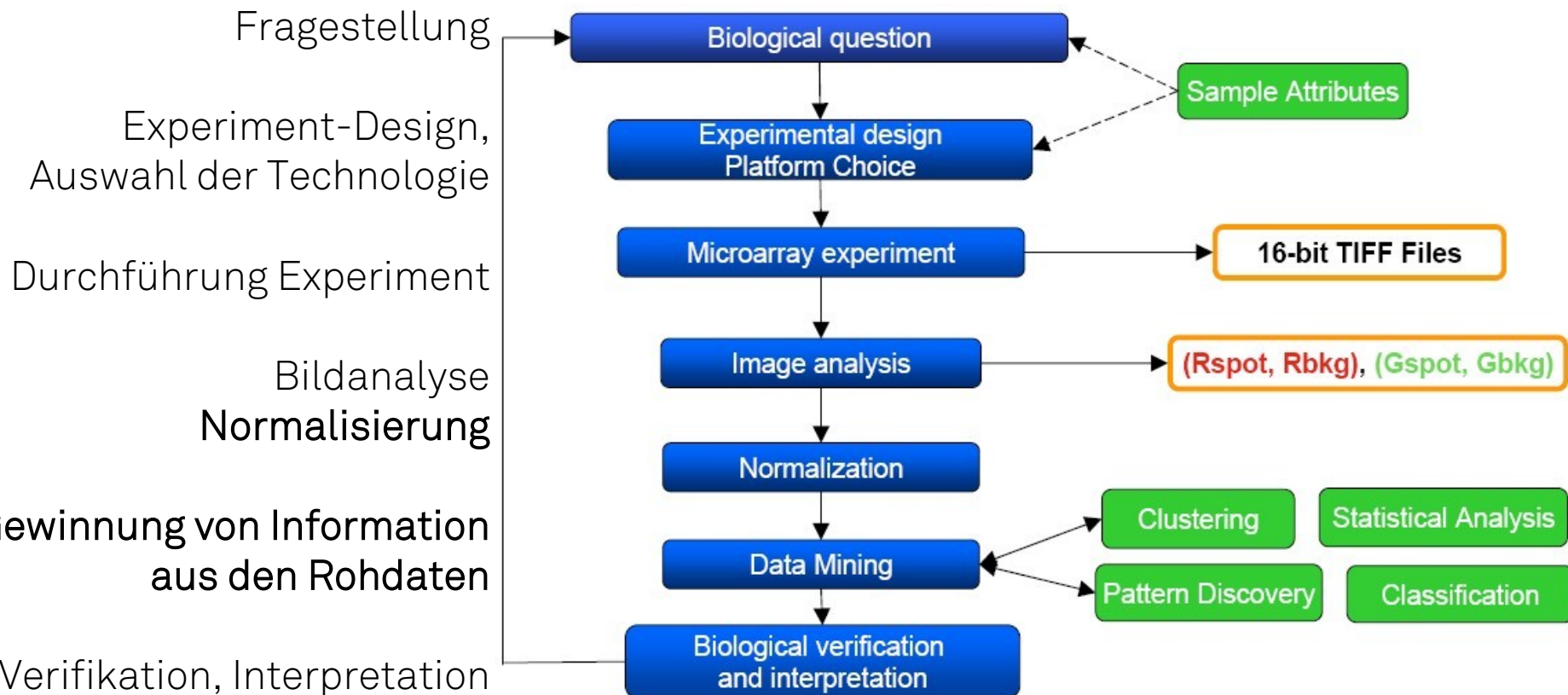
Amplifikation und Markierung der RNA



Fragmentierung  
Hybridisierung

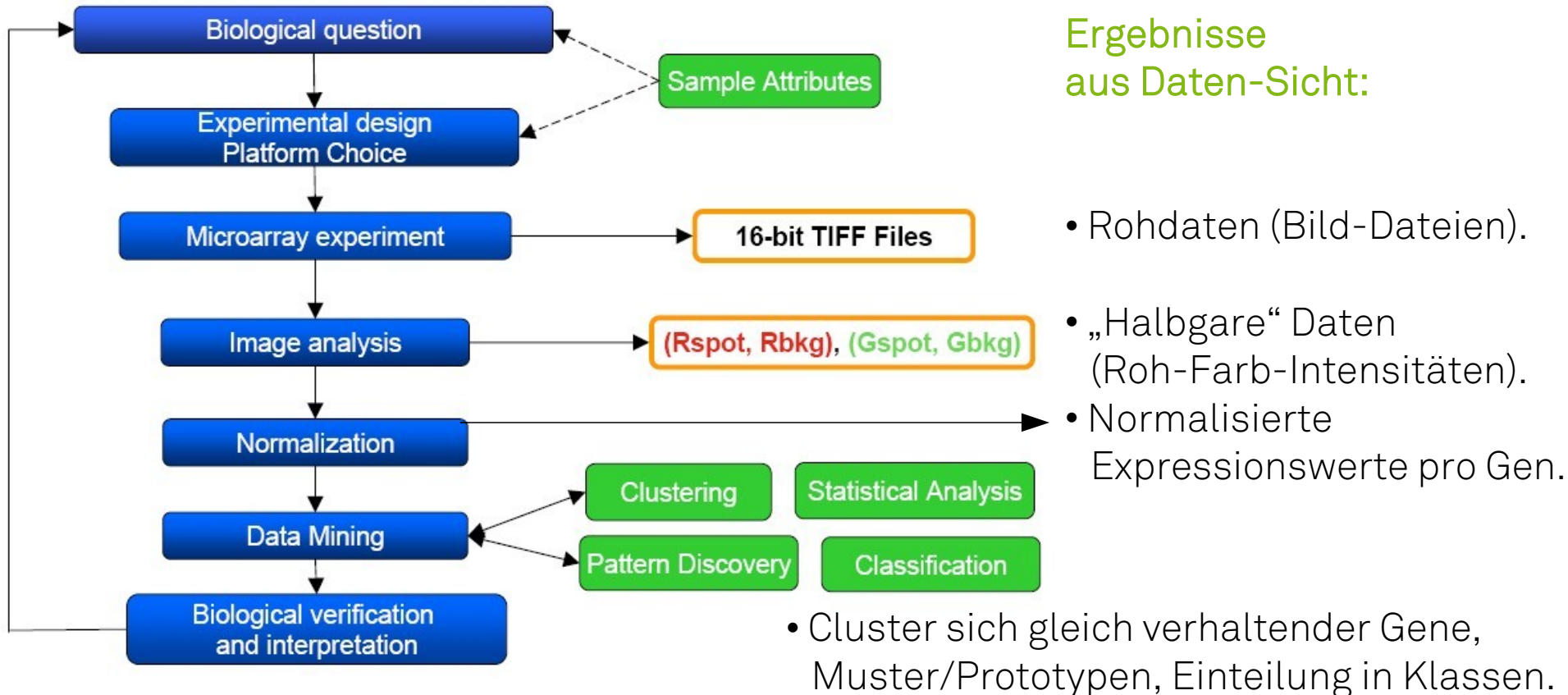


## Schritte einer Microarray-Studie





## Resultate einer Microarray-Studie



### Ergebnisse aus Daten-Sicht:

- Rohdaten (Bild-Dateien).
- „Halbgare“ Daten (Roh-Farb-Intensitäten).
- Normalisierte Expressionswerte pro Gen.

- Cluster sich gleich verhaltender Gene, Muster/Prototypen, Einteilung in Klassen.

### Ergebnisse aus Wissens-Sicht:

- Unterstützung oder Falsifikation der initialen Hypothese; neue Hypothese(n).
- wenn keine Hypothese vorlag: Generierung von Hypothesen.

## Auswertung einer Microarray-Studie

Typische Visualisierung einer Microarray-Studie:

### hierarchisches (Bi-)Clustering

- Zeilen: Gene
- Spalten: Experimente (z.B. Zeitpunkte)

**Rot:** Gen unterexprimiert gegenüber Standard.

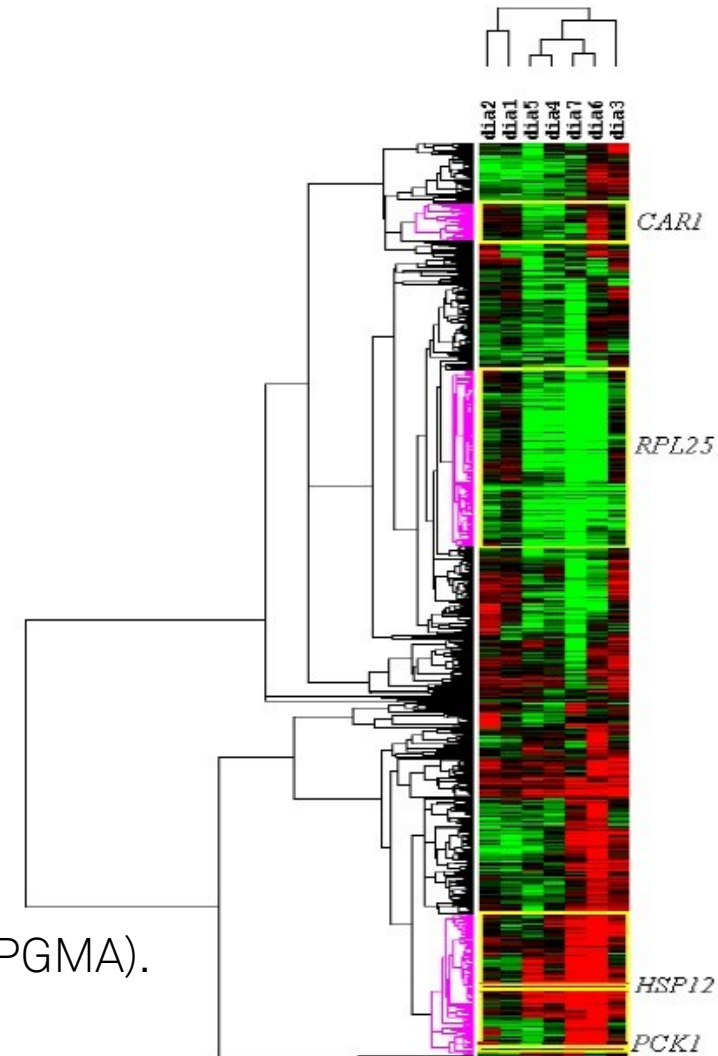
**Grün:** Gen überexprimiert gegenüber Standard.

Gene, die sich über viele Experimente gleich verhalten, bilden Gen-Cluster  
Experimente, in denen sich viele Gene gleich verhalten, bilden Experiment-Cluster.

Anordnung der Gene und Experimente:

Ähnliche möglichst nebeneinander.

Berechnung wie bei phylogenetischen Bäumen (UPGMA).



## Die einzelnen Analyseschritte

### Low-Level-Analyse:

- Bildanalyse
- Hintergrundkorrektur
- Normalisierung (Herausrechnen von systematischen Fehlern)
- Schätzung der absoluten oder relativen Genexpression pro Gen

### High-Level-Analyse:

- Identifizierung differenziell exprimierter Gene
- Clustering und Klassifikation von Genen und Experimenten

## Low-Level: Bildanalyse und Hintergrundkorrektur

### Eingabe:

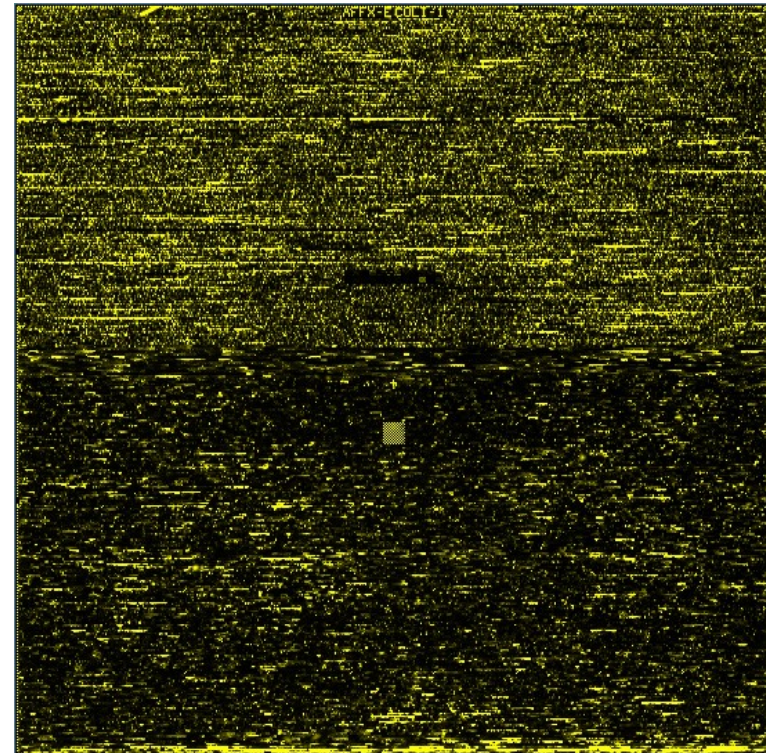
- fotografierte Bilddaten (ein Intensitätswert pro Bildpunkt)

### Vorgehen:

- Identifikation der Lage der Spots
- Berechnung einer Gesamtintensität pro Spot aus Bildpunkten, die Spot bilden
- Abzug des „Hintergrunds“
- Transformation (z.B. Logarithmierung)

### Ausgabe:

- ein Intensitätswert pro Spot/Feature



## Low-Level: Normalisierung

### Problem:

- Intensitätswerte pro Spot nicht direkt vergleichbar für mehrere Arrays
- Grund z.B.: mehr Farbstoff; längere Belichtung auf zweitem Array
- Daher z.B. alle Werte auf Array 2 doppelt so groß wie auf Array 1, aber Gene nicht doppelt so aktiv.

### Grundannahme der Normalisierung:

Zwischen zwei Experimenten (Arrays)

- ändert sich die Expression einzelner Gene
- ändern sich nicht globale Eigenschaften der Verteilung der Expressionswerte

Daher:

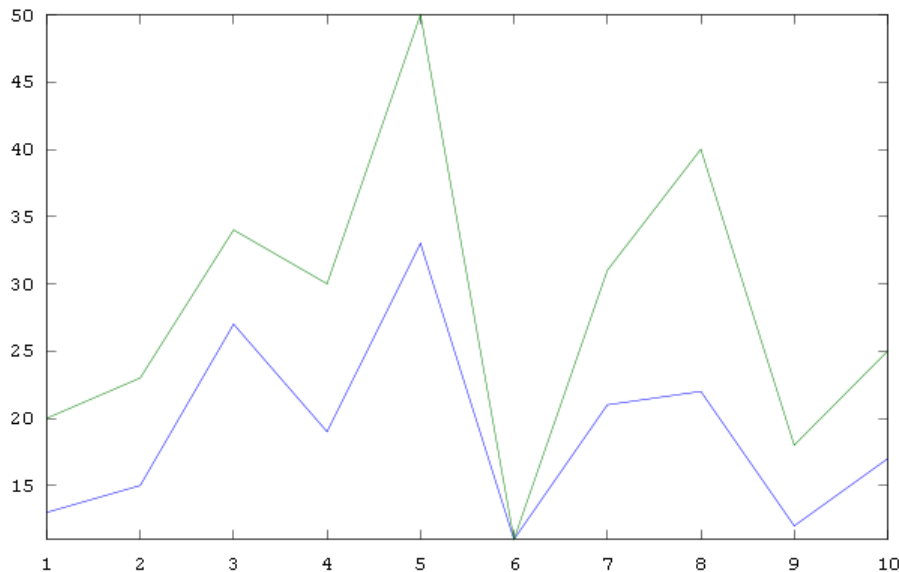
Angleichung der Histogramme möglich und sinnvoll.

## Beispiel zur Normalisierung

Zwei „Mini“-Microarrays mit nur 10 Spots:

Array 1 in blau: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17)

Array 2 in grün: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25)



Alle Werte in Array 2 scheinen höher.

Widerspricht Grundannahme, also: angleichen = normalisieren.

## Affin-Lineare Normalisierung

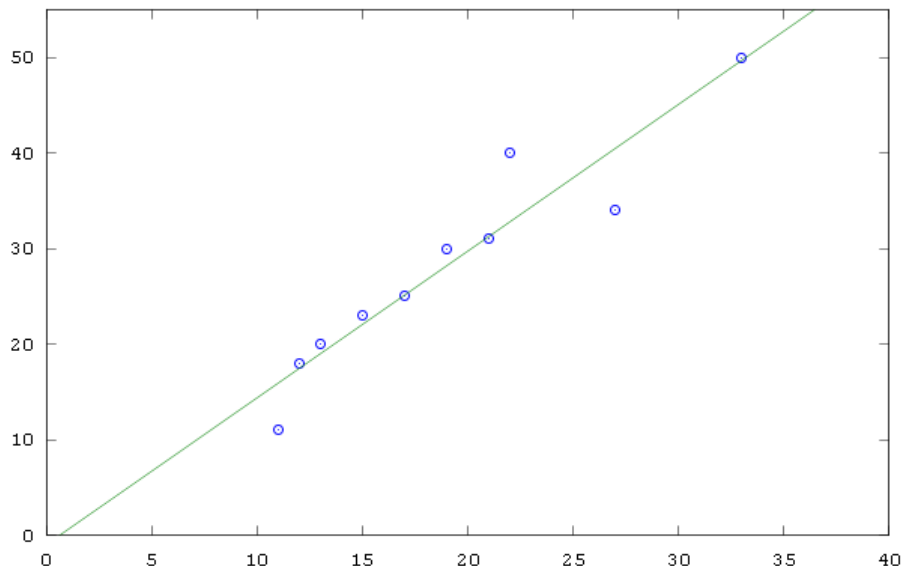
### Annahme:

Es gilt ein affiner Zusammenhang  $y = ax + b$

mit  $x$ : Wert aus Array 1,  $y$ : entsprechender Wert aus Array 2.

Scatterplot von  $x$  gegen  $y$  (blau) und

Berechnung einer Ausgleichsgeraden (grün):  $y = 1.53394x - 0.94480$ .





## Affin-Lineare Normalisierung

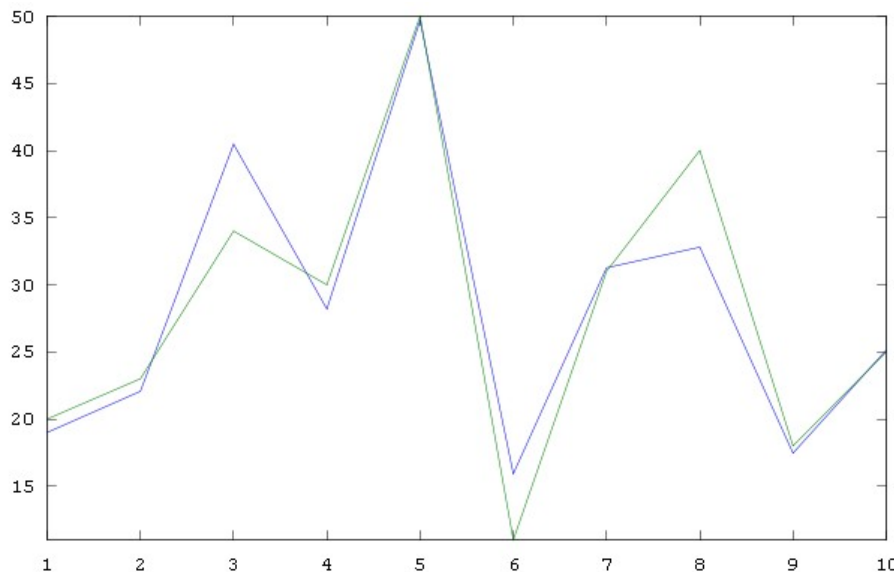
Zwei „Mini“-Microarrays mit nur 10 Spots:

Array 1: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17)

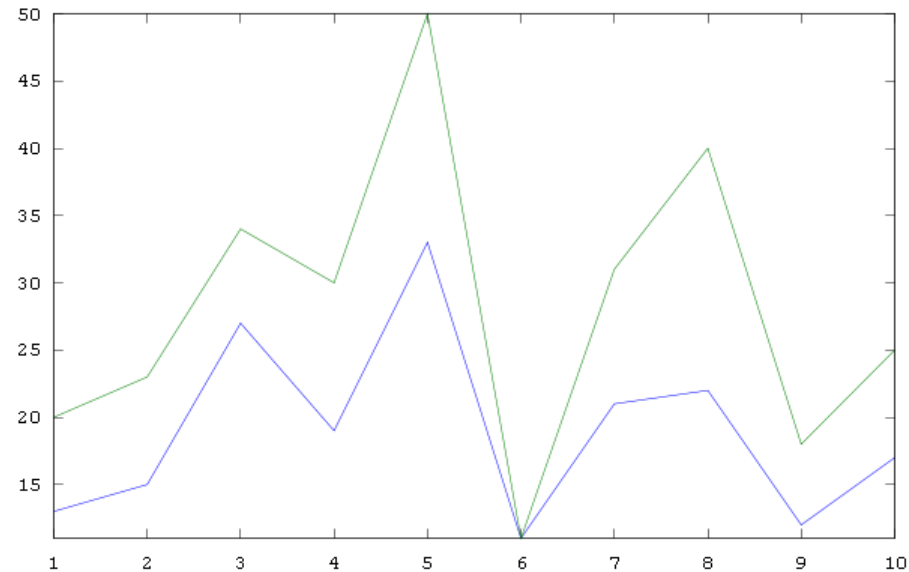
angleichen mittels  $y = 1.53394x - 0.94480$ .

Array 1' in blau: (19.0, 22.0, 40.5, 28.2, 49.7, 15.9, 31.3, 32.8, 17.5, 25.1)

Array 2 in grün: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25)



(nach Normalisierung)



(vor Normalisierung)



## Quantil-Normalisierung (RMA-Verfahren)

**Annahme:** Sortiert man die Intensitäten beider Arrays, sind die Werte gleich.  
D.h. Alle Quantile der Intensitätsverteilungen sind gleich (und gleich Referenz).

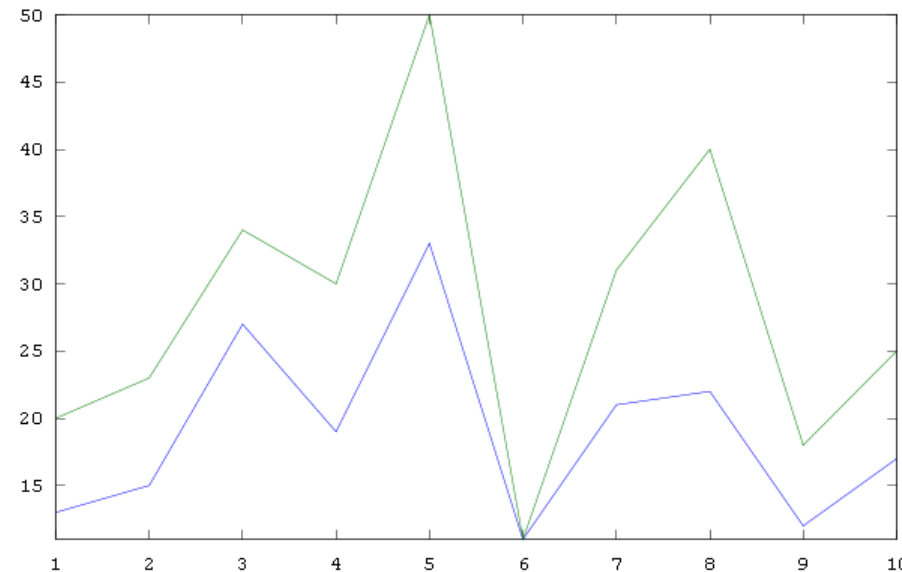
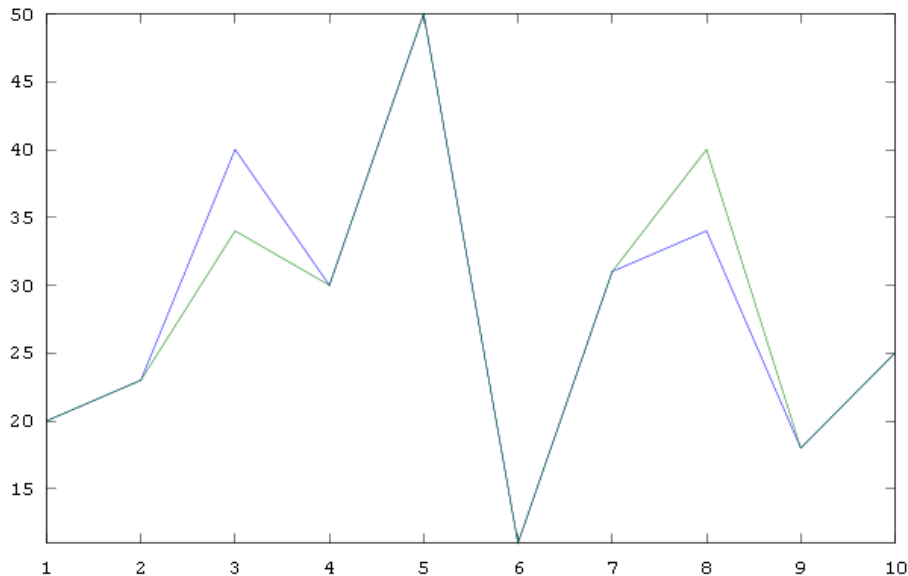
Array 1: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17) # x

Array 2: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25) # y

Array 1': (20, 23, 40, 30, 50, 11, 31, 34, 18, 25) # x normalisiert

Array 1' enthält die Werte aus Array 2, aber Sortierung aus Array 1.

**in R:** `x[order(x)] = referenz` # referenz: sortierte y-Werte



## Schätzung eines Expressionswerts pro Gen

Jedes Gen wird durch mehrere (z.B. 11) Sonden / Spots gemessen.  
Idealerweise liefern alle nach Normalisierung denselben Wert.

Das ist aber in der Praxis nicht so.

Die 11 Intensitätswerte werden zu einem Expressionswert zusammengefasst.

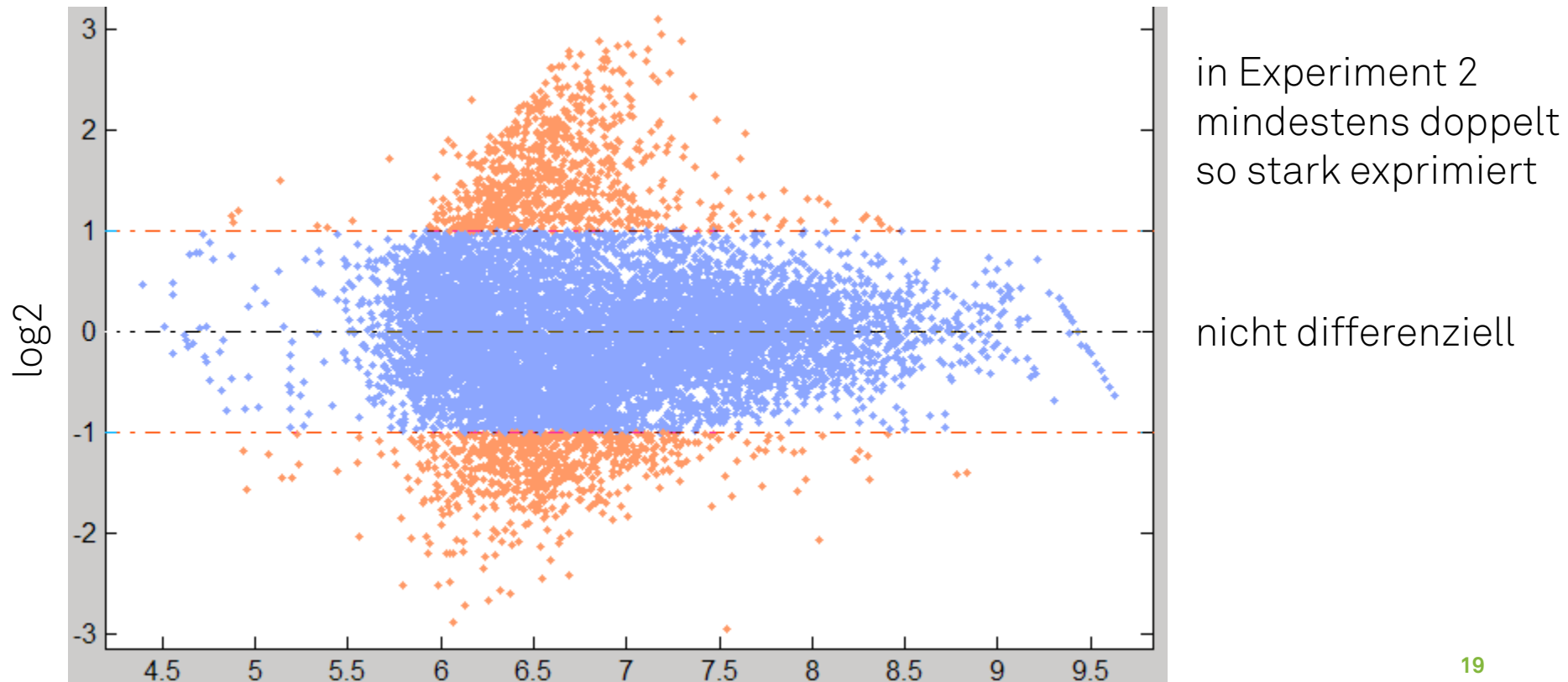
### Methoden:

- Mittelwert
- Median
- Modelle, die die Hybridisierungsstärke der Sonden berücksichtigen.

## Identifikation differenziell exprimierter Gene (Kandidaten-Gene)

MA-Plot („Minus gegen Average“):

Differenz gegen Durchschnitt aller Genexpressionswerte in zwei Experimenten.



## MIAME - Standard

### MIAME: Minimal Information About a Microarray Experiment

Richtlinien der MGED (Microarray and Gene Expression Data) Society

URL: [http://www.mged.org/Workgroups/MIAME/miame\\_2.0.html](http://www.mged.org/Workgroups/MIAME/miame_2.0.html)

Ziel: Informationen zu Microarray Studien in Datenbanken sollen langfristig nutzbar und interpretierbar bleiben.

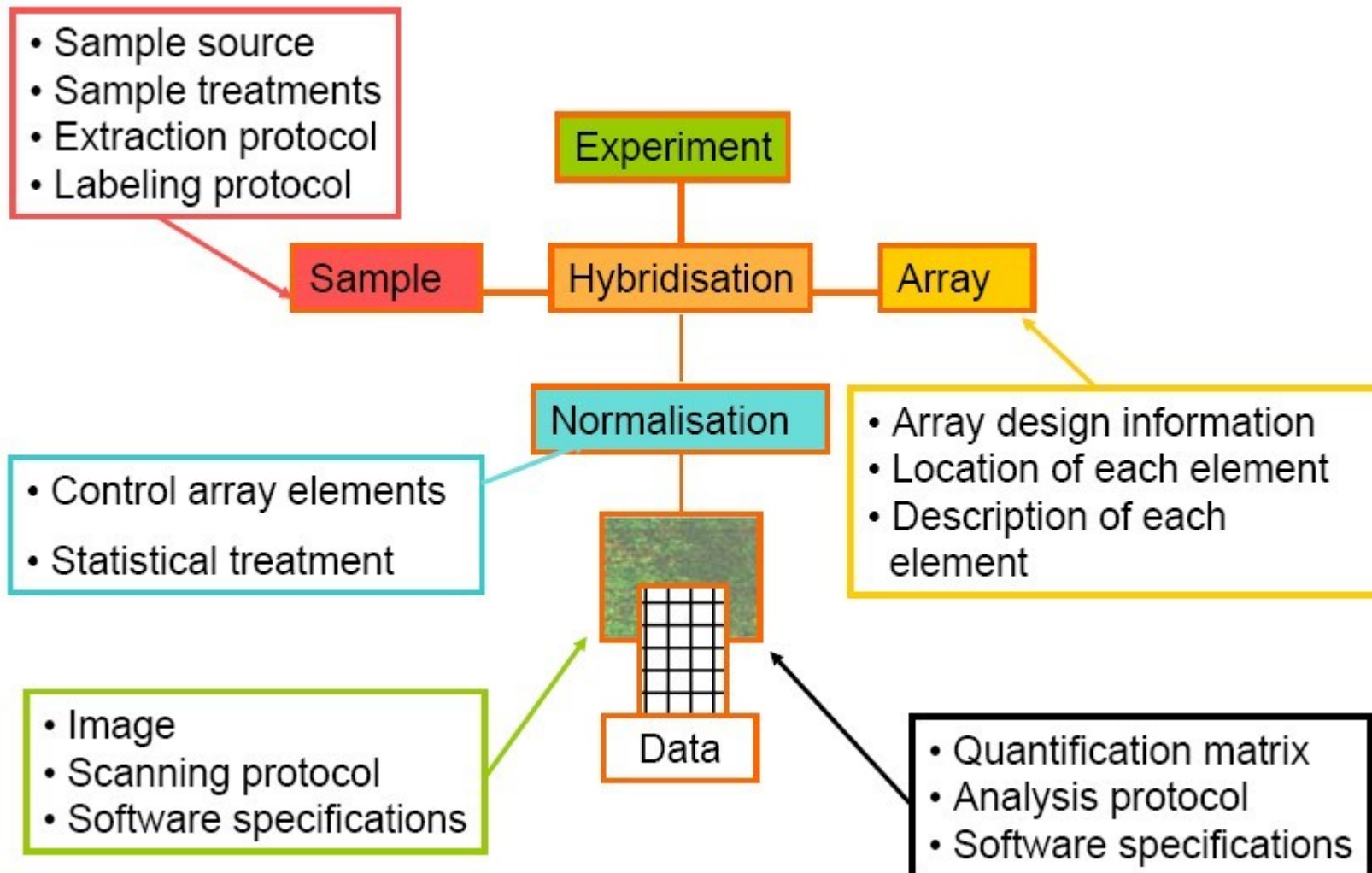
Verlangt werden Informationen zu 6 Bereichen.

## Nach MIAME verlangte Informationen

### 6 Bereiche:

1. Rohdaten (Bild-Daten, unbearbeitet).
2. Prozessierte Daten (Matrix mit Expressionswerten pro Gen und Experiment).
3. Daten zu den Proben (samples):  
Woher stammen die Proben (z.B. Tumorgewebe / gesundes Gewebe);  
wie wurden in jedem Experiment die Proben behandelt?
4. Daten zum experimentellen Design:  
Welche Beziehungen bestehen zwischen den Experimenten einer Studie,  
zwischen verschiedenen samples ?  
Welche Rohdaten gehören zu welchem sample?
5. Daten zum Array-Design:  
Entweder welches Array von welchem Hersteller,  
oder bei Eigenentwicklungen Liste der DNA-Sequenzen aller Sonden.
6. Protokolle, sowohl experimentell, als auch zur Datenanalyse.

## Nach MIAME verlangte Informationen



# GEO

## Gene Expression Omnibus

- MIAME-konforme Microarray-Datenbank
- am NCBI
- <http://www.ncbi.nlm.nih.gov/geo/>

The screenshot shows the NCBI Gene Expression Omnibus (GEO) website. At the top, there is the NCBI logo and the GEO logo with the text "Gene Expression Omnibus". Navigation links include HOME, SEARCH, SITE MAP, Handout, NAR 2006 Paper, NAR 2002 Paper, FAQ, MIAME, and Email GEO. The current page is "NCBI > GEO" and the user is "Not logged in".

The main content area features a description of the Gene Expression Omnibus: "a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval."

There are three main navigation sections:

- QUERY**: Includes links for DataSets, Gene profiles, GEO accession, and GEO BLAST, each with a search input field and a "GO" button.
- BROWSE**: Includes links for DataSets, GEO accessions, Platforms, Samples, and Series.
- SUBMIT**: Includes links for Direct deposit / update, Web deposit / update, and Create new account.

On the right side, there are two summary boxes:

- Public data**: Shows counts for GPL Platforms (4754), GSM Samples (231841), GSE Series (8901), and Total (245496).
- Site contents**: Lists various resources under "Documentation" (Overview, FAQ, Submission guide, etc.) and "Query & Browse" (Repository browser, Submitter contacts, etc.).

At the bottom, there is a search bar for "Get GEO accession" with fields for Scope (Self), Format (HTML), and Amount (Quick), along with a "GO" button. Below that is a "Depositors only" section with fields for User and Password, a "LOGIN" button, and a link to "Recover a password".


The footer contains links for | NLM | NIH | GEO Help | Disclaimer | Section 508 |

## ArrayExpress

- MIAME-konforme Microarray-Datenbank
- am EBI
- <http://www.ebi.ac.uk/microarray/>

### Microarray Informatics at the EBI

#### ArrayExpress - a public resource for transcriptomics and related data

<b>Query ArrayExpress:</b>	Profiles ▾	<input type="text" value="nfkbia"/>	Gene	
<a href="#">Browse ArrayExpress</a>		<input type="text" value="leukemia"/>	Keyword	
		All ▾	Species	
				<input type="button" value="Query"/>

[Submit data to ArrayExpress](#)

[Analyse data in Expression Profiler](#)

**New** [ArrayExpress Atlas Beta](#)

[Documentation](#)

Defined MGED Standards: [MIAME](#) [MAGE-TAB](#) [MAGE-ML](#) [MGED Ontology](#)



## Ein GEO-Beispiel

Accession Number GSM120719

URL: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM120719>

Standard Affymetrix Experiment; Homo sapiens; Muskelgewebe.

Prozessierte Daten, ein Intensitätswert pro Transkript: einfache Textdatei.

3 Spalten: Transkript-ID; Intensitätswert; Present/Absent - Entscheidung

ID_REF	VALUE	ABS_CALL
200000_s_at	7844.43408203125	P
200001_at	31178.318359375	P
200002_at	65324.34375	P
200003_s_at	60674.9921875	P
200004_at	28636.283203125	P
200005_at	16733.904296875	P
200006_at	39892.4765625	P
200007_at	23791.220703125	P
200008_s_at	10948.0546875	P
200009_at	18589.80078125	P
200010_at	48697.41015625	P
200011_s_at	3656.510009765625	P

...

## Software für die Analyse von Microarray-Experimenten

### Anbieter von Analyse-Software

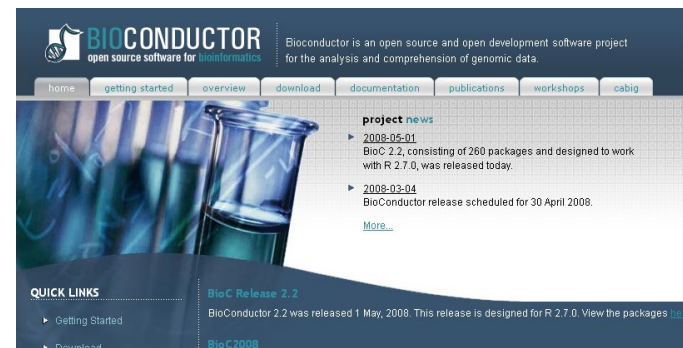
- Gerätehersteller
- Kommerzielle Anbieter (spezialisierte Firmen)
- Forschergruppen, oft freie Software

### Was ist Freie Software?

- „free as in speech“
- „free as in beer“

Bioconductor für R: Paketsammlung, Freie Software.

- URL: <http://www.bioconductor.org/>
- u.a. Funktionen für Microarrays



## Zwischenspiel: Was macht ein/e Bioinformatiker/in ? Beispiel Microarray-Design

### Vorbemerkung

Es gibt verschiedene „Arten“ von Bioinformatikern.

Was hier steht, gilt für mich und ähnlich ausgebildete Bioinformatiker.

- Diplom in Mathematik
- Promotion in (algorithmisch orientierter) Informatik
- wissenschaftliche Arbeit in der Algorithmik

## Was macht ein/e (algorithmische/r) Bioinformatiker/in ?

### Wichtig zu wissen

In der bioinformatischen Forschung werden eher nicht existierende Programme angewendet, sondern eher neue Ideen (und auch Programme) entwickelt. Man kann aber nur etwas Neues entwickeln, wenn man weiß, was es schon gibt!

### Serviceleistungen

- Implementierung dieser Algorithmen („Programmieren“)
  - Anwenden des Programme auf konkrete Daten
- sind nicht Teil der wissenschaftlichen Forschung in der Bioinformatik, aber wichtig in der Praxis und für Kooperationen mit anderen Fächern!

## Was macht ein/e (algorithmische/r) Bioinformatiker/in ??

### Modellbildung / Modellierung

- Übersetzen des Problems in ein Modellproblem (Formalisierung)
- Fortwährende Verbesserung der Modellierung (Realitätsnähe)

### Algorithmik

- Entwurf von Algorithmen, die das formale Problem korrekt und schnell lösen
- Fortwährende Verbesserung der Algorithmen (Effizienz)

## Typischer Arbeits-Prozess

Man hört oder liest von einem Problem in den Lebenswissenschaften,  
das bisher nicht zufriedenstellend gelöst ist,  
z.B.: Welche Sonden (Oligonukleotide) sollen auf einem Microarray sein?

- Man sammelt Informationen darüber:
  - Wer leidet unter dieser Situation? [„Marktanalyse“]
  - Mit welchen Methoden wurde es versucht? [Literaturrecherche]
- Man versucht, eine bessere Idee zu haben als alle anderen ... [Kreativität]
- Man probiert die Ideen aus (und modifiziert sie, ggf. mehrere Male) [Forschung]
- Man zeigt, dass die neu erfundenen Ansätze besser/schneller sind als bisherige.
- Man erstellt eine (einfach benutzbare) Software, die diese Ideen enthält.
- Man schreibt die Publikation(en) darüber.
- Man trägt über seine Resultate vor [Marketing].

## Problemstellung: Design von Microarrays

### Gegeben

- Genomsequenz (Annahme: vollständig sequenziert und korrekt assembliert)
- Menge von Transkripten (mRNA-Sequenzen)

### Gesucht

- spezifische Oligonukleotid-Sonden für jedes Transkript

### Modellierung von „spezifisch“:

Der „Auftraggeber“ will gute Ergebnisse bei seinen Experimenten.

Die molekulare Dynamik der Hybridisierungsreaktion ist nicht 100%ig verstanden.

Der Bioinformatiker muss aus dem, was gewünscht ist, mit Hilfe von Fachwissen ein einfaches Modell erstellen, mit dem er arbeiten kann.

Zur Verfügung stehen Sequenzen.

## Modellierung

### Erste Idee:

Ein Oligonukleotid  $s$  ist spezifisch für ein Transkript  $t$ , wenn  $s$  in  $t$  als Teilsequenz vorkommt, aber sonst nirgendwo im Genom.

### Kritik:

Das ist zu grob modelliert! Auch zu beachten:

- reverse Komplemente der Sequenzen,
- mögliche Selbst-Hybridisierung des Oligonukleotids,
- Schmelztemperatur,
- Position des Oligos relativ im Transkript
- ...



## Formalisierung

### Gegeben

- Genomsequenz (String)  $s$
- Menge von Transkript-Sequenzen  $T = \{t_1, t_2, \dots, t_n\}$ , jedes  $t_i$  ist Teilstring von  $s$
- Oligonukleotid-Länge  $L$

### Gesucht

- Für jedes Transkript  $t_i$ :  
Teilstrings der Länge  $L$ , die in  $t_i$  und nirgendwo sonst in  $s$  vorkommen

### Erklärungen

- Ein **String** (Zeichenkette) ist dasselbe wie eine Sequenz.
- Ein **Teilstring** ist ein (zusammenhängender) Ausschnitt aus einem String.
- Das Zeichen  $\leftarrow$  bedeutet: „wird zugewiesen an“.

## (Ineffizienter) Algorithmus zur Auswahl spezifischer Sonden

Für jedes Transkript  $t$ :

    Für jeden Teilstring  $u$  der Länge  $L$  von  $t$ :

        anzahl[ $u$ ]  $\leftarrow$  0

    Für jeden Teilstring der Länge  $L$  von  $s$ :

        Wenn  $u=v$ : anzahl[ $u$ ]  $\leftarrow$  anzahl[ $u$ ] + 1

    Wenn anzahl[ $u$ ]=1: gib  $u$  als Kandidaten für  $t$  aus

### Laufzeit

$|T| * |s| * L$  Vergleiche,

Humangenom: 33 M \* 3 G \* 25,

also  $2.5 * 10^{18}$  Vergleiche,

bei  $10^9$  pro Sekunde sind das ca. 80 Jahre!

## Diskussion

Welche Fragestellungen kennen Sie aus Ihrem Studium,  
aus Ihrer wissenschaftlichen Arbeit (Bachelorarbeit),  
bei denen der Einsatz von Informatik-Methoden nützlich ist / sein könnte?

Versuchen Sie, die Ausgangssituation (was ist gegeben?)  
und das gewünschte Ergebnis möglichst genau zu beschreiben.