

Molekularbiologische Datenbanken

Genexpression
Analyse und Datenmanagement

Ulf Leser

Wissensmanagement in der
Bioinformatik



mRNA - cDNA

5' GpppGCAUCGCAUUAUGCGAAGGGCUUUGCAUUGAaaaaaaaaaaaaaaaaaaaaa 3'

(dGTP)_n

(dCTP)_n

(dATP)_n

(dTTP)_n



- Trick: Rück-Übersetzung von mRNA in cDNA
 - Reverse Transscriptase (RT)
 - Primer z.B. am Poly-A Tale (oder zufällig)
- Clonierung der cDNA in Libraries

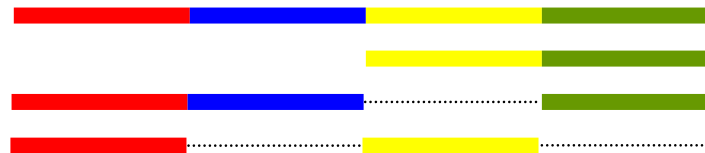
EST Clustering

- 4.5 Millionen ESTs – wie viele Gene ?
- Clustern der EST nach Sequenzüberlappungen
- Ähnlich Assembly, aber
 - Hohe Fehlerrate in Sequenz
 - Differential Splicing
 - ESTs sind auf Exons zusammengesetzt
 - Überlappungen liegen in „Blöcken“ vor
 - Löcher dazwischen (gespleisste Exons) stören nicht
- Verwendung spezieller Algorithmen
- Vorsicht beim Vergleich mit genomischer DNA: **keine Introns**

Original Gen
(ohne Introns)



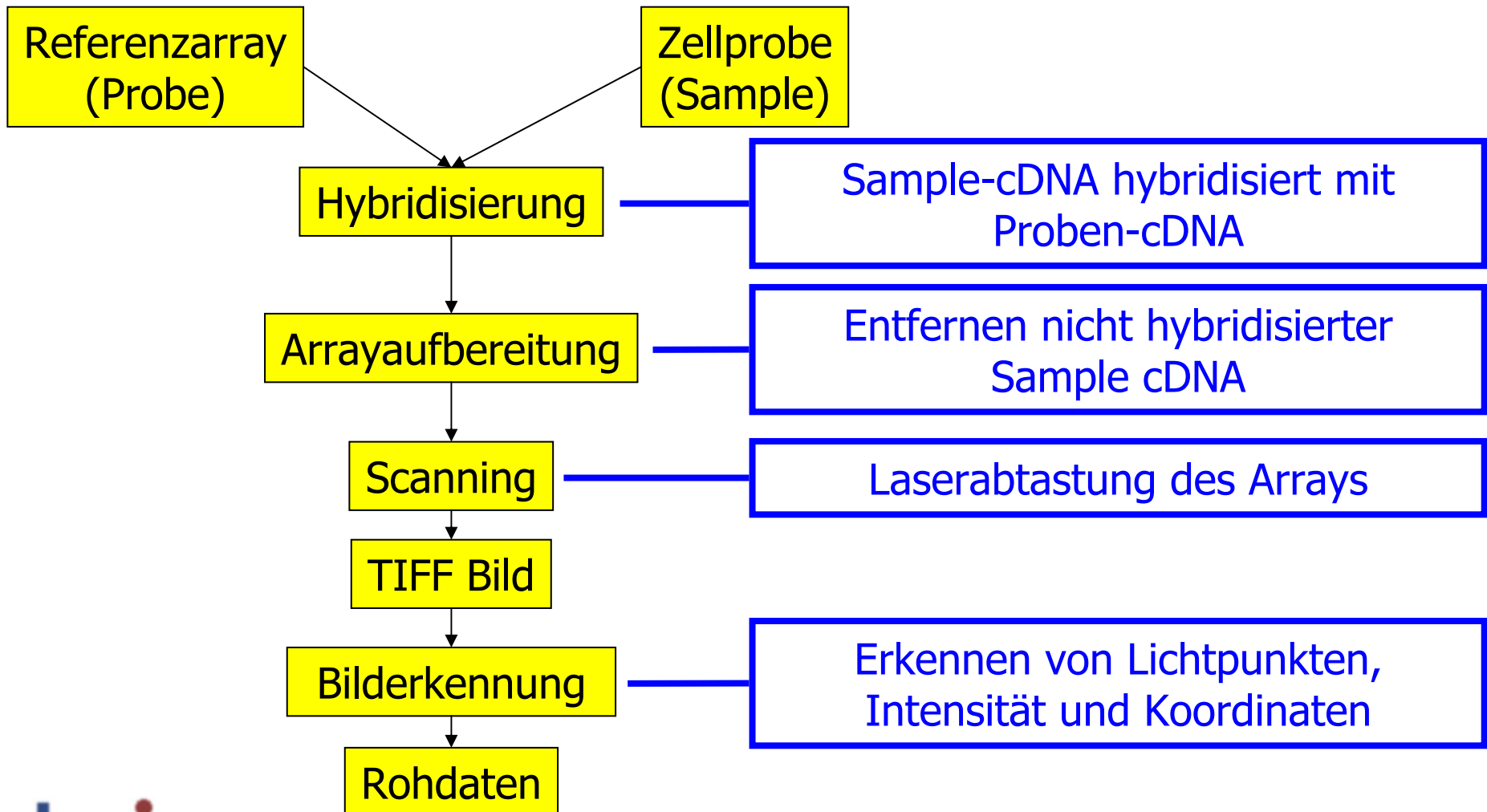
Mögliche EST Varianten



Microarrays - Grundidee

- Wir haben jetzt cDNA Bibliotheken
- Durch Clustering kann man auch redundanzarme Bibliotheken erstellen
- Was kann man damit machen?
 - Microarrays, Genchips, Genexpressionmessung
 - Bibliothek alles cDNA (z.B: UniGene): Probe
 - cDNAs aus Zellen extrahieren: Sample
 - cDNAs in Sample hybridisieren mit cDNAs als Probe
 - Messen der Expressionsniveaus aller Gene einer bestimmten Zelle zu einem bestimmten Zeitpunkt

Ablauf



Probleme

- **Gesund–krank Messung schwierig**
 - Normale Expression gibt es i.d.R. nicht
 - Isolierung von Zellen u.U. schwierig
- **Genrepräsentation**
 - Viele Gene nur selten und in geringer Dosis aktiv (insbesondere embryonale Zellen)
 - Geringe Menge in Samples – nicht nachweisbares Signal
- **RNA Isolierung**
 - Jede Zellmanipulation zur Verarbeitung induziert Veränderung in Genexpression (Stress, Apoptose, ...)
- **Normalisierung**
 - Vergleich Ergebnisse über Laborgrenzen hinweg sehr schwierig

Inhalt dieser Vorlesung

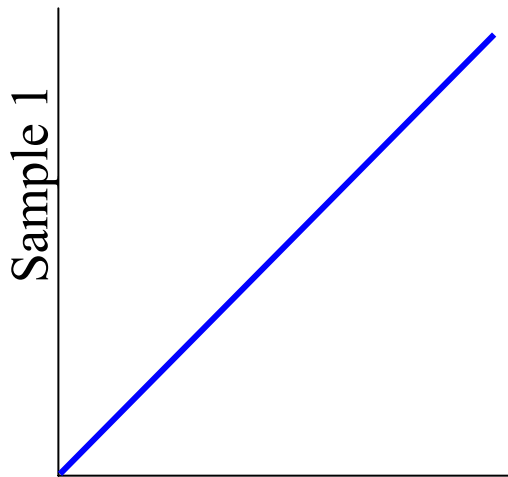
- Differentielle Expression
- Clustering zur Ko-Regulation
- Datenmanagement und MIAME

Rohdaten

- **OligoArrays**
 - Expressionsintensitäten einzelner Gene
 - Experimentreihen: X Kontrollen, Y Krank
- **cDNA Arrays**
 - Expressionsratios einzelner Gene
 - Experimentreihen: X (Kontrolle, Krank)
- **Im folgenden: OligoArrays**

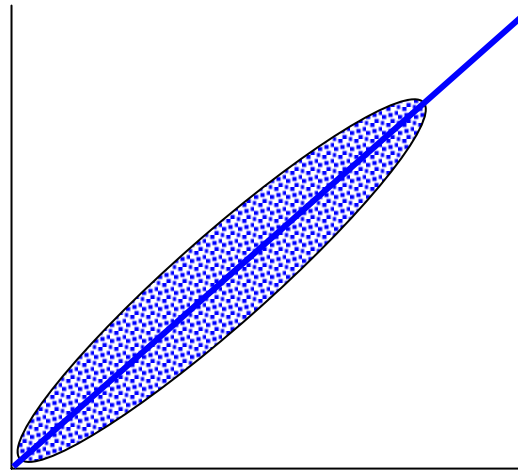
Visualisierung - Scatterplots

Ein Punkt = Ein Gen



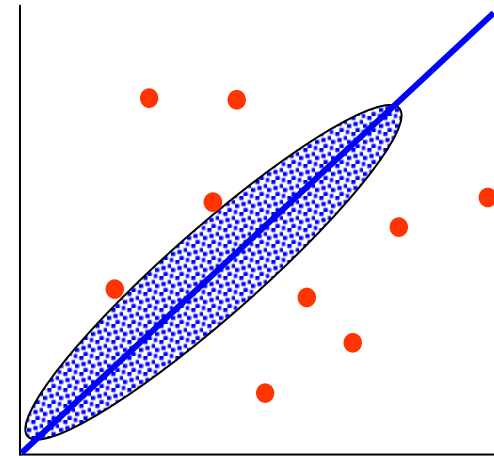
Sample 2

Vollständig
identische
Exprimierung



Sample 2

Verteilung von
Intensitätsun-
terschieden

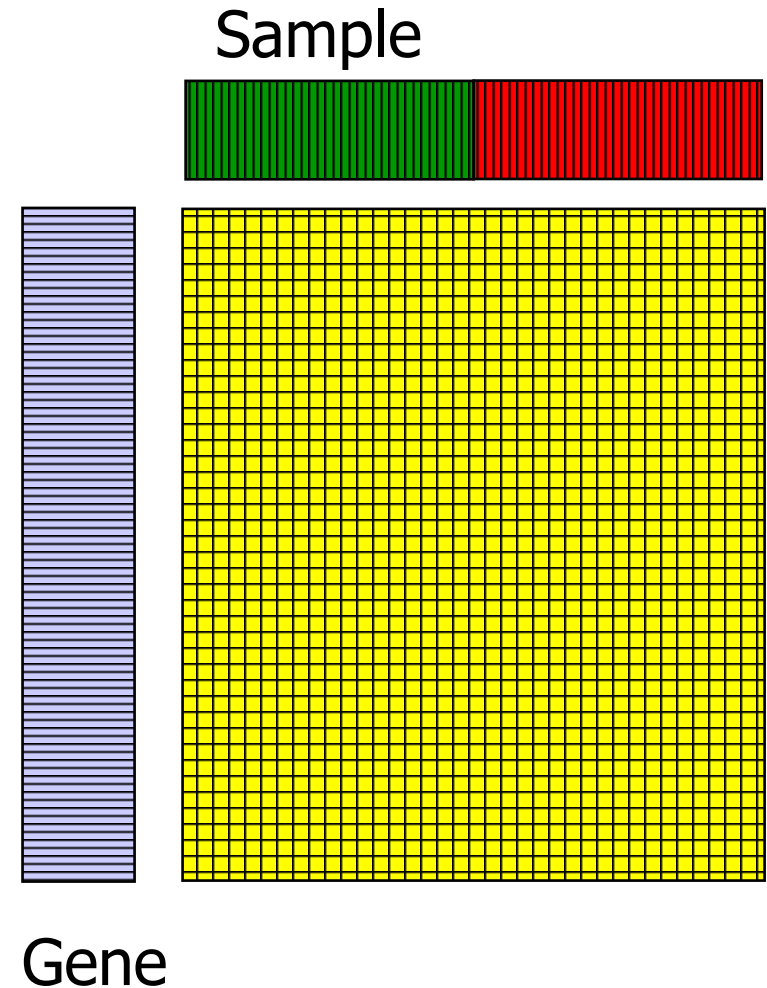


Sample 2

Ausreißer:
Interessante
Gene

Differentielle Expressierung

- Annahme
 - S_1, \dots, S_m : Gesunde Sample
 - T_1, \dots, T_n : Kranke Sample
- Gesucht: Gene mit **signifikanten** Unterschieden zwischen S und T
- Werte eines Gen X
 - $S = \{s_1, \dots, s_m\}$
 - $T = \{t_1, \dots, t_n\}$
- Zwei Verfahren
 - Simple Fold
 - T Test



Methode 1: Simple Fold

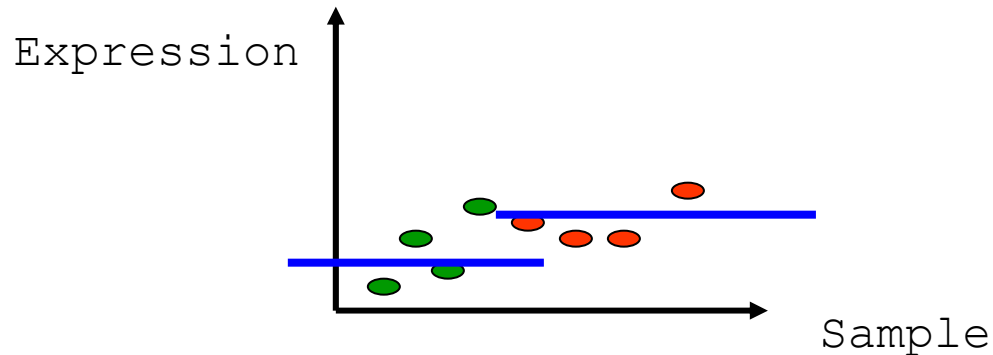
– X differentiell exprimiert, gdw.:

$$e^{\left| \log \left(\frac{\text{avg}(T)}{\text{avg}(S)} \right) \right|} > t$$

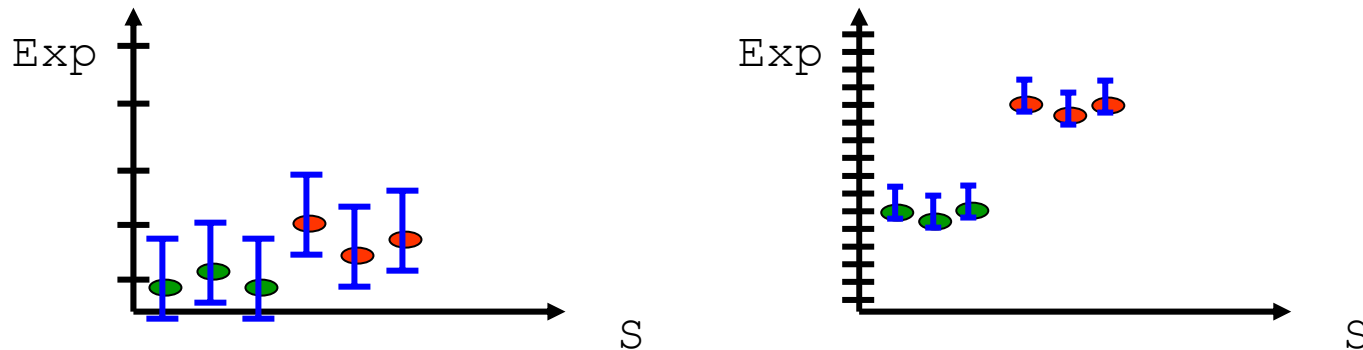
- log: Gleichbehandlung von Steigerung/Verringerung
- Signifikanz wird definiert durch Schwellwerte t, z.B.
 - <2: Uninteressante Veränderung
 - 2-4: interessant
 - >4: sehr interessant

Probleme des Simple Fold

- Vergleicht nur zwei Werten – die Mittelwerte

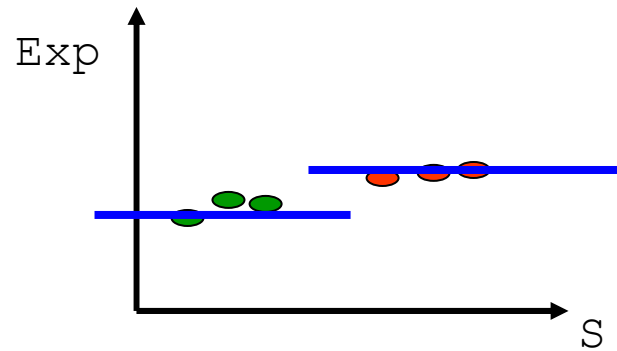
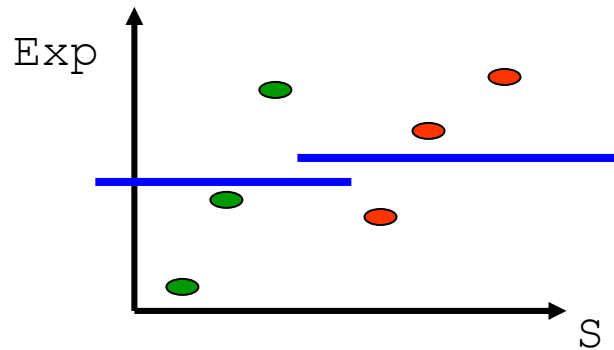


- Unabhängig von absoluten Größen und Fehlerraten



Probleme 2

- Unabhängig von Streuung



Statistischer Test: t-Test

- T-Test

- Aussage über die **Signifikanz de Unterschieds** zwischen den Werten einer Testreihe und einer Gesamtheit
- Signifikanzniveau α
 - Wahrscheinlichkeit für ein falsch negativ vorhergesagtes Ergebnis des t-Tests
- Beispiel
 - Herstellung von Folien mit Dicke 0,25 cm
 - Dicke folgt Normalverteilung (Mittelwert = 0.25 cm)
 - Testreihe: 10 Folien, Mittelwert 0.253, SD 0.003
 - Frage: Arbeitet die Maschine korrekt mit Sicherheit α ?

- Annahmen

- **Normalverteilung** der Werte
- Kleine Stichprobe (< 30)
 - Sonst werden andere, aber ähnliche Tests verwendet

Anwendung des t-Test

- Expressionsniveaus „Gesund“ ist Gesamtheit (S)
- Expressionsniveaus „Krank“ ist Testreihe (T)
- T-Test Wert

$$t = \frac{avg(S) - avg(T)}{\sqrt{\frac{sd(S)^2}{m} + \frac{sd(T)^2}{n}}}$$

- t: Stärke der differentiellen Expressierung
- X differentiell exprimiert mit Signifikanz α gdw. $|t| < \text{STUDENT}(\alpha)$
- $\text{STUDENT}(\alpha)$: Erlaubte Abweichung nach Verteilungstabelle

Beispiel

	s ₁	s ₂	s ₃	s ₄	t ₁	t ₂	t ₃	t ₄
Gen X	9	11	10	8	20	22	21	15
Gen Y	1	2	1	2	2	2	4	4
Gen Z	9	11	10	8	12	14	10	10

	X	Y	Z
Simple Fold	1.37	1.35	1.07
T Wert	-5.14	-2.01	-1.36

- Signifikanz der Aussage abhängig von gewünschter false-negative Rate
- P-Value: kleinstes α , für das T-Test noch signifikant

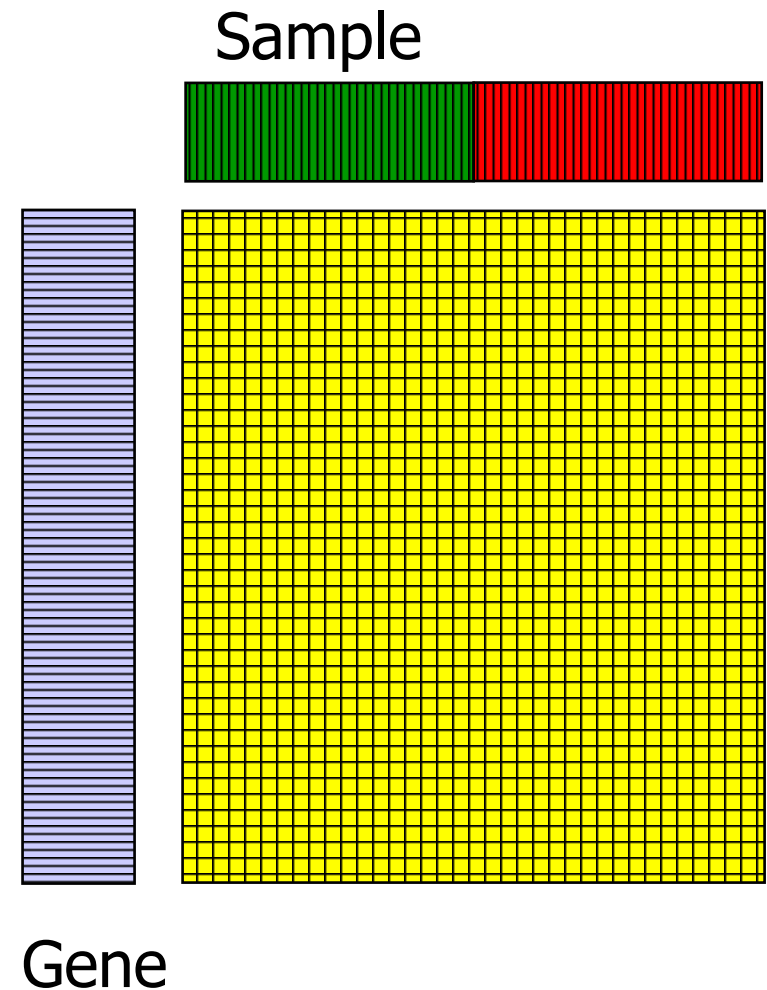
Differentielle Expressierung

- Simple Fold und t-Test verbreite Methoden
 - t-Test beachtet Mittelwerte und **Varianzen**
 - Probleme bei kleinen n,m
- Vorsicht
 - 10.000 Gene, $\alpha = 0.01 \rightarrow 100$ falsch negative Ergebnisse
- Weitere Methoden
 - Probabilistische Modellierung: Berechnung der Wahrscheinlichkeit der beobachteten Werte unter Annahme einer bestimmten Werteverteilung
 - Regressionmodelle

-
- Bisher: Erkennen des auffälligen Verhaltens eines Gens
 - Jetzt: Erkennen, welche Gene gemeinsam auf einen Stimulus reagieren
 - Z.B. auf Gabe eines Medikamentes

Ko-Regulation

- Annahme
 - S_1, \dots, S_m : Gesunde Sample
 - T_1, \dots, T_n : Kranke Sample
- Gruppen von Genen mit gleichartiger Veränderung zwischen S und T
- Vergleich und Clustering von Zeilen
 - Hierarchisch
 - K-means



Ko-Regulation

- Finden von Genen mit „gleicher“ Veränderung
 - Welche Gene reagieren gemeinsam auf Temperaturstress?
 - Gruppierung nicht bekannt
 - Clustering – „Unsupervised learning“
 - Methoden: Hierarchisch, K-Means, Principal Comp. Analysis, ...
- Finden von Gruppen von charakteristischen Genen
 - Welches Expressionsmuster ist für Tumorart X charakteristisch
 - Robuster als Untersuchung nur einzelner Gene
 - Beispiele für Gruppen bekannt (gesund – krank)
 - Klassifizierung – „Supervised learning“
 - Methoden: Bayes'sche Classifier, Support Vector Machines
 - Hier nicht weiter behandelt – Data Mining

Voraussetzung

- Zeitreihen

- Keine Trennung von Gesund / Krank mehr
- Gen X = $\{s_{x1}, \dots, s_{xn}\}$
- Gen Y = $\{s_{y1}, \dots, s_{yn}\}$
- Punkte im n-dimensionalen Raum

- Ähnlichkeitsmaße zwischen zwei Genen

- Euklidischer Abstand $\sqrt{(s_{x1} - s_{y1})^2 + \dots + (s_{xn} - s_{yn})^2}$
- Korrelationskoeffizient (Winkel der Vektoren)
- ...

- Cluster: Gruppen mit

- Hoher Ähnlichkeit zwischen Mitgliedern
- Geringer Ähnlichkeit zu Nicht-Mitgliedern

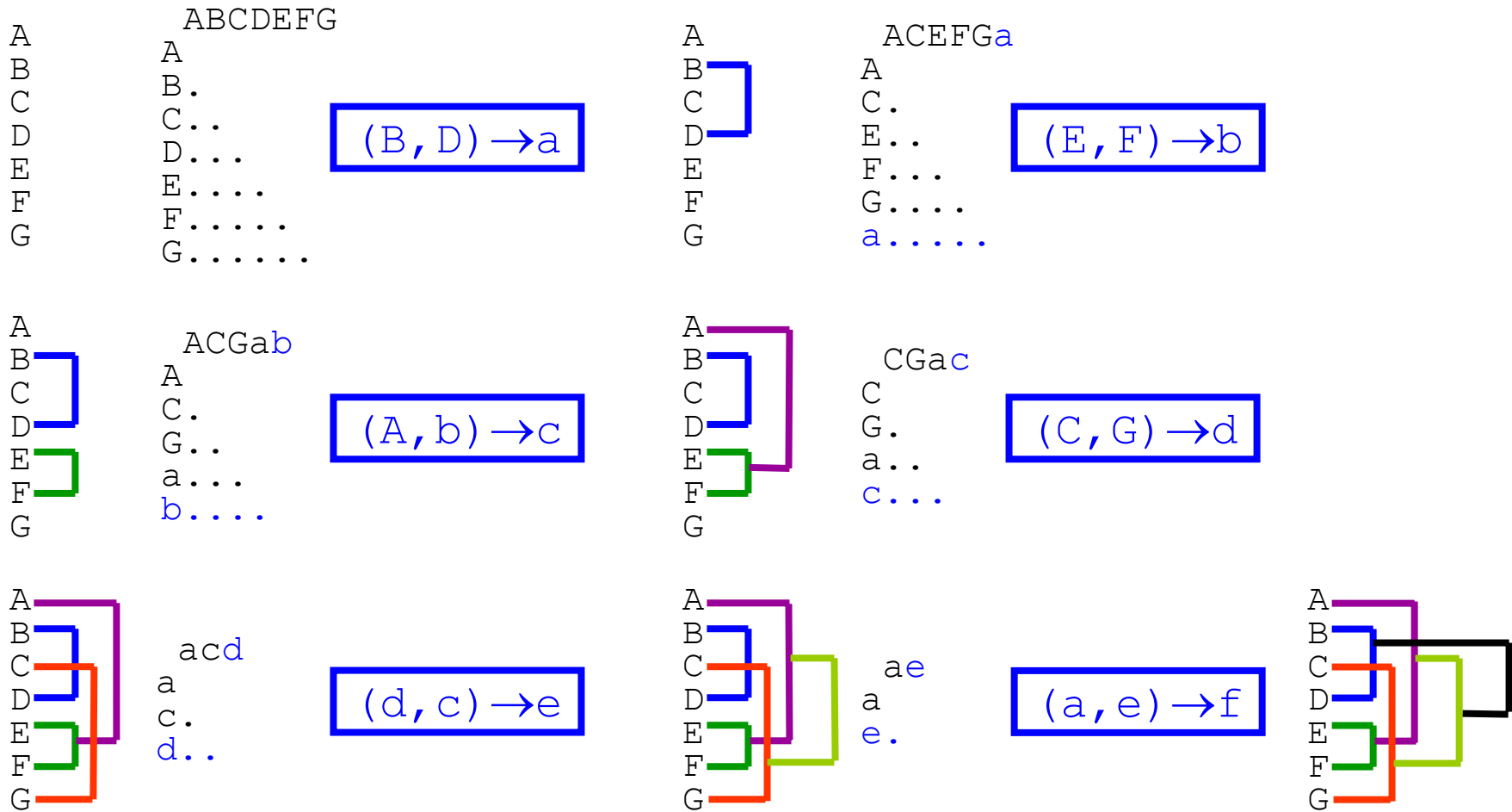
Hierarchisches Clustering

- Sukzessive Berechnung ähnlichster Paare
 - Hierarchische Visualisierung der Paare
- Vorteile
 - Sehr einfaches Verfahren
 - Effiziente Algorithmen
 - Intuitive Visualisierung
 - Hoher Bekanntheitsgrad
- Nachteile
 - Im eigentlichen Sinne keine Clusterbildung
 - Wenig Hilfestellung bei Interpretation

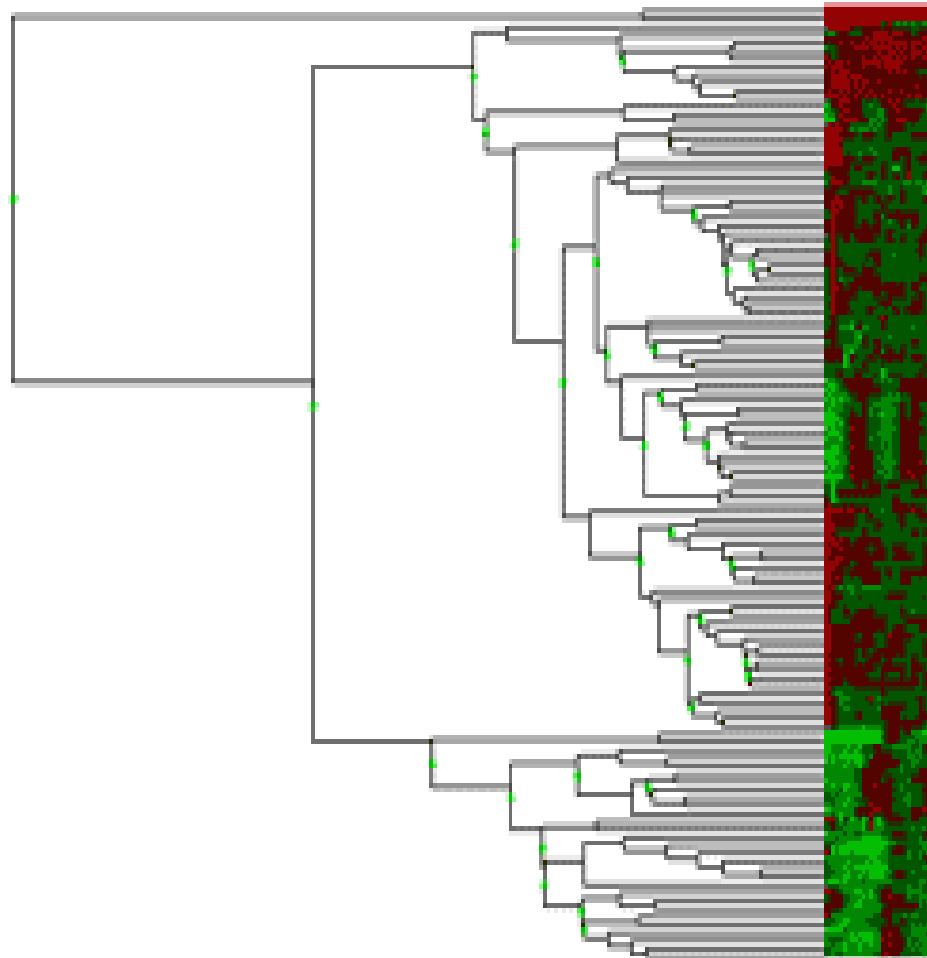
Algorithmus

1. Berechnung der Ähnlichkeitsmatrix A
2. Solange $|A| > 1$
 1. Auswahl des ähnlichsten Paares (X,Y)
 2. Berechnung von $Z = \text{avg}(X,Y)$ $\left\{ \frac{s_{x1} + s_{y1}}{2}, \dots, \frac{s_{xn} + s_{yn}}{2} \right\}$
 3. Löschen von X und Y in A
 4. Einfügen von Z in A
 5. Berechnung des Abstandes von Z zu allen Elementen von A
 6. Visualisieren / notieren von X und Y als Paar

Graphisch



Reale Daten



Quelle: <http://www.ii.uib.no/~bjarted/jexpress/hclust.html>

Komplexität

- Berechnung der Ähnlichkeitsmatrix

$$\sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}$$

- Berechnung der Ähnlichkeiten zu Z in Schritt k

– Pro Schritt: $n - k - 1$, insgesamt: $\sum_{i=1}^{n-2} (n - i - 1) = \frac{(n-1)(n-2)}{2}$

- Zusammen: $O(n^2)$

Bewertung

- Ordnung der Gene zu „konfliktarmen“ Graphen:
 - 2^n Ordnungen
 - $O(n^4)$ Algorithmus bekannt
- Ergebnis ist binärer Baum
 - Ableitung von Clustern bleibt Benutzern überlassen (Schwellen für Ähnlichkeit)
- Keine Aussage über Güte des Clusterings
- Variante: Länge der „Arme“ von Ähnlichkeit der Knoten abhängig machen

K-Means - Voraussetzungen

- Menge von Genen: $X = \{X_1, X_2, \dots, X_n\}$
- Festlegung der Menge an Clustern $k < n$
- Zufällige Auswahl von k Genen $Y_1, \dots, Y_k \in X$
 - Das sind die initialen Clusterzentren
- Festlegung eines Schwellwerts s
- Hilfskonstruktionen
 - $\text{centre}[C_j]$: Clusterzentrum von C_j
 - $\text{cluster}[X_i]$: Zuweisung von X_i zu Cluster
 - $\text{distance}(X_i, C_j)$: Abstand von X_i zu $\text{center}[C_j]$
 - $\text{elements}(C_j)$: Menge aller Gene in Cluster C_j

K-Means - Algorithmus

- $\forall j: \text{center}[C_j] = Y_j$
- $\forall i: \text{cluster}[X_i] = -1$
- $t = \text{MAXINT}$

Erste Clusterzentren zufällig gewählt

Initial keine Zuordnung

- Solange $t > s$

Solange mehr als s Bewegungen

– $t = 0$

– Für alle $X_i \in X$

Berechne Clusterzuordnung aller Gene

- $\text{Old} = \text{cluster}[X_i]$
- $j = \min[j] \text{ distance}(X_i, C_j)$
- $\text{cluster}[X_i] = j$
- if $\text{old} \neq j: t=t+1$

Hat Gen X_i seine Clusterzuordnung geändert?

– Für alle C_j

- $\text{centre}(C_j) = \text{avg}(\text{elements}(C_j))$

Berechne neue Clusterzentren

K-means Bewertung

- Komplexität?
- Verschiebt Gene solange, bis Cluster konvergieren
 - Konvergenz nicht gesichert
- Clustergüte ableitbar
 - Mittlerer Abstand jedes Gens zum Zentrum seines Cluster
- Menge an Clustern muss festgelegt werden
 - „Gute Zahl“ bedingt gute Clustergüten
 - Mit mehreren Werten experimentieren
- Schwellwert s
 - zu hoch: schlechte Cluster
 - zu niedrig: lange Laufzeit, Gefahr des Oszillierens
- Startzentren beeinflussen Ergebnis massiv
 - Mit mehreren Anfangskonstellationen durchführen

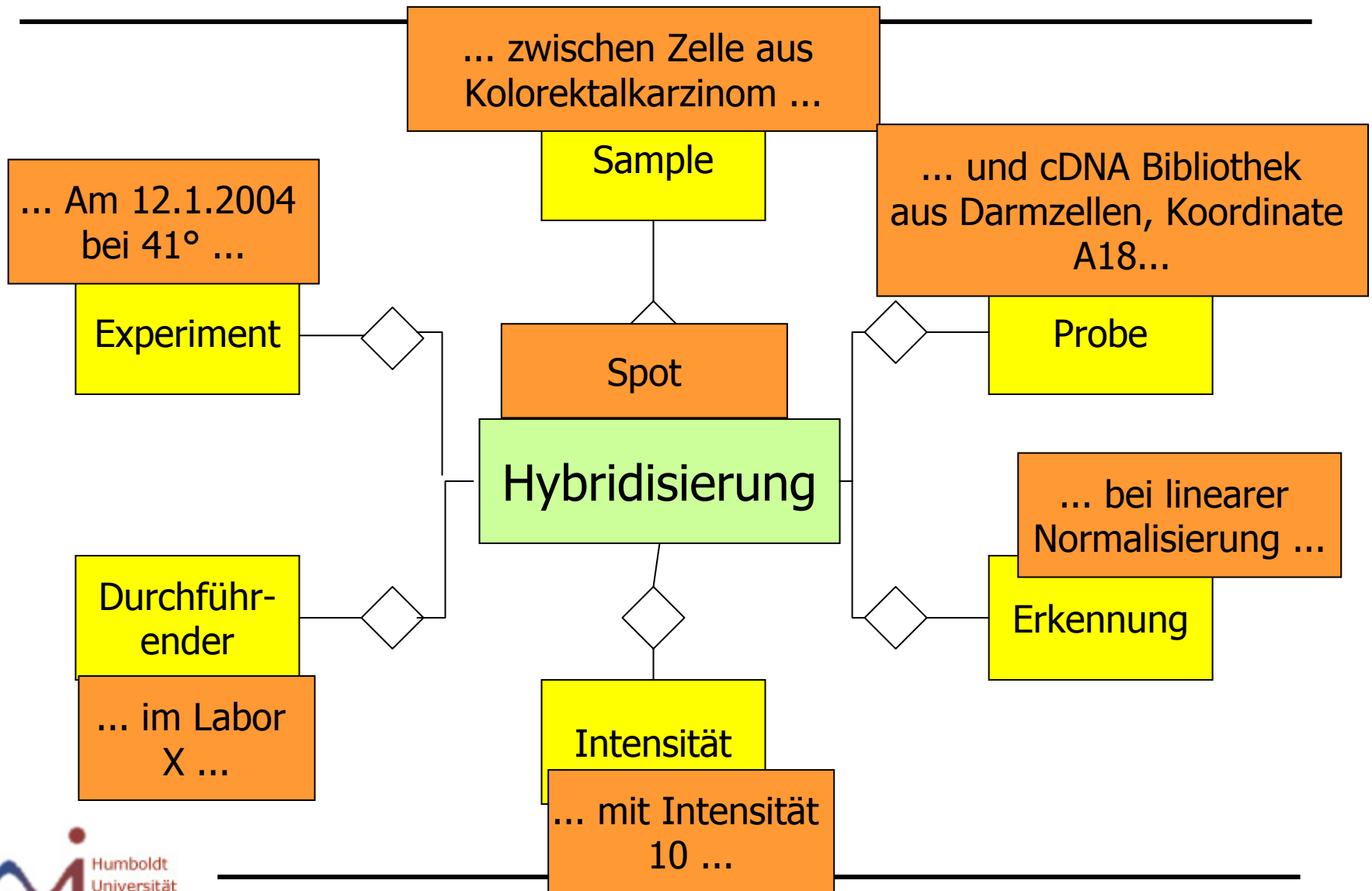
Zusammenfassung

- Hauptkomponenten der Datenanalyse
 - Differentielle Expression: Signifikanz von Änderungen
 - Simple Fold
 - T-Test
 - Ko-Regulation von Genen: Gemeinsame Aufgaben
 - Hierarchisches Clustering
 - K-Means
- Viele Methoden und Implementierungen (EXCEL, R, ...)
- Viele kommerzielle und Open Source Produkte verfügbar
- **Unterschiede sind erheblich**
 - 20 (T-Test) oder 200 (Regression) interessante Gene [Pa02]

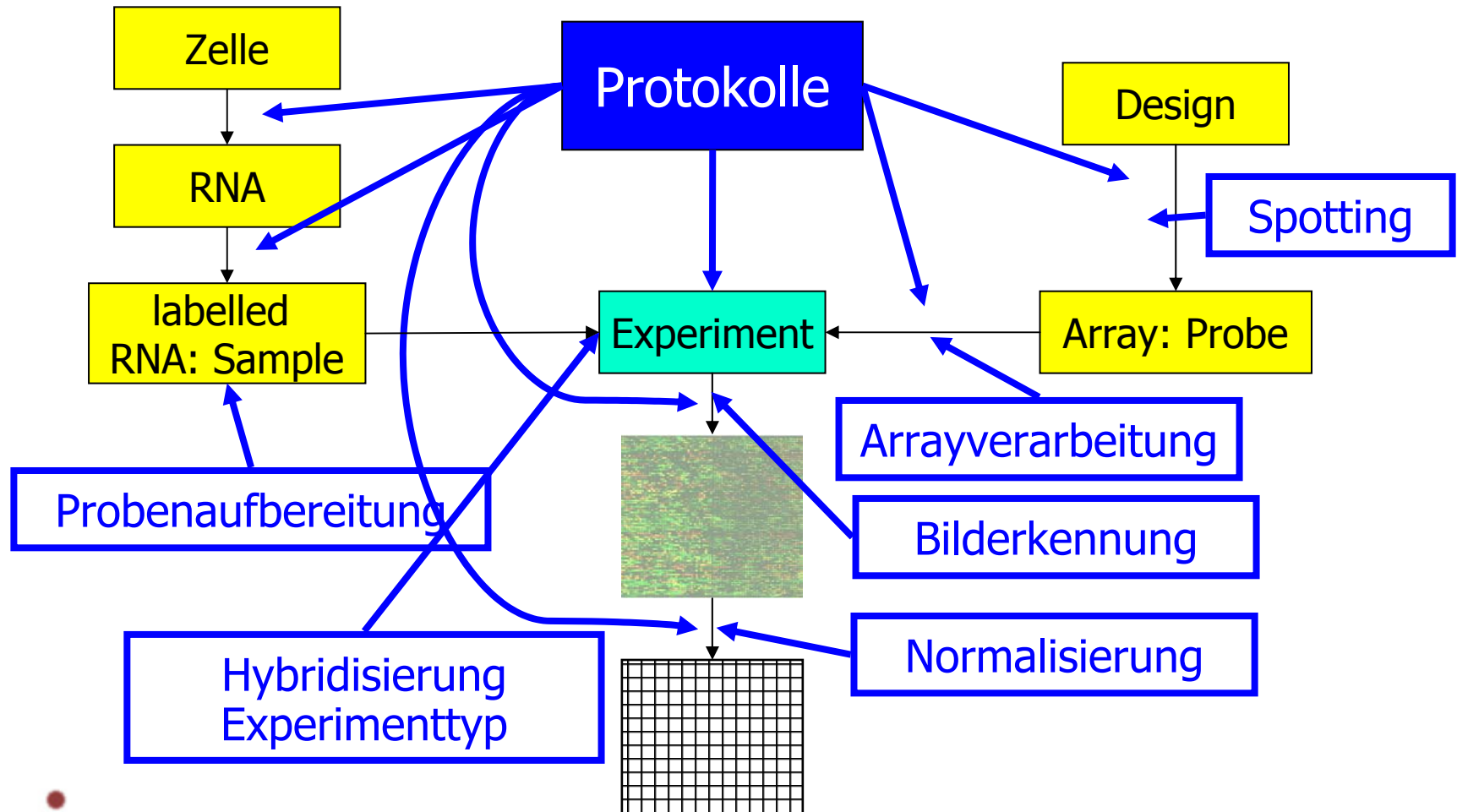
Teil III. Datenmodelle

- MAGE Objektmodell und MAGE ML

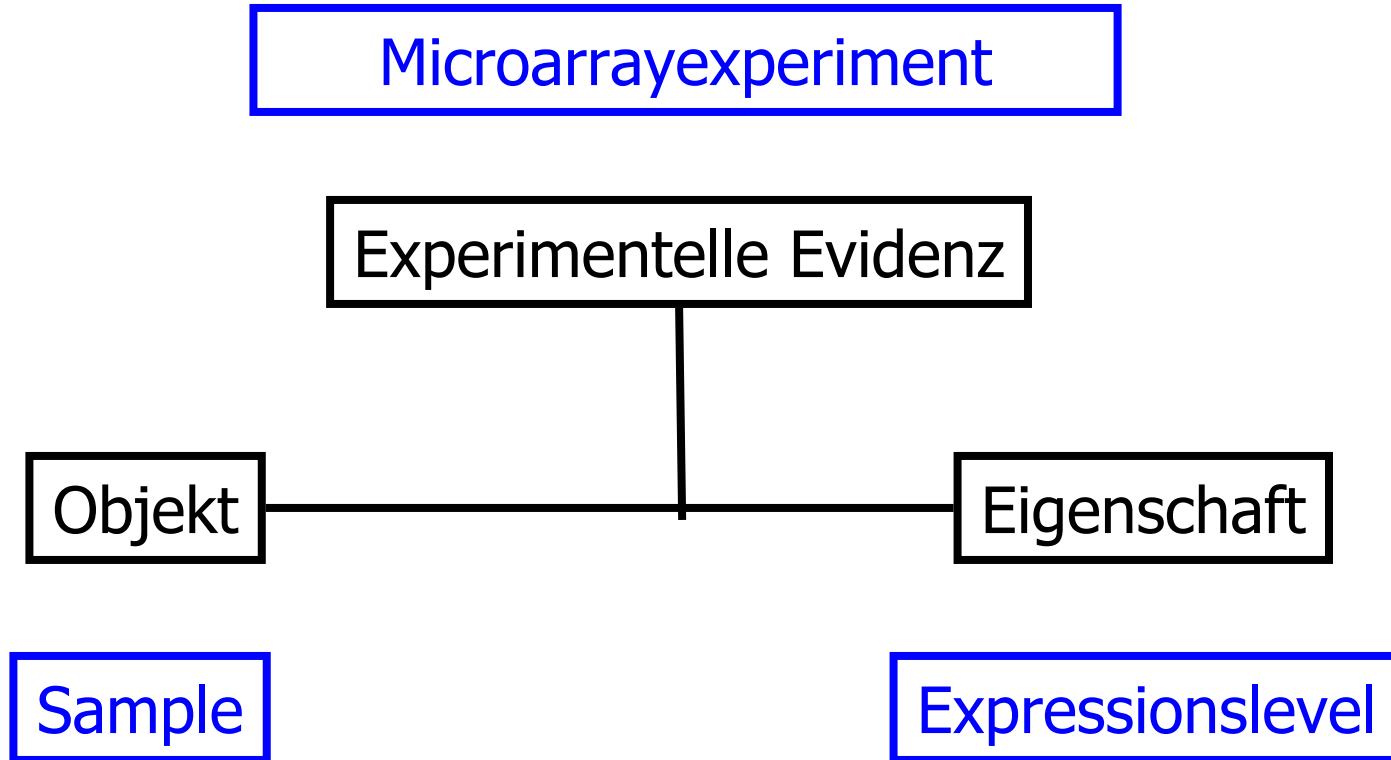
Genexpression - Daten



Prozesse



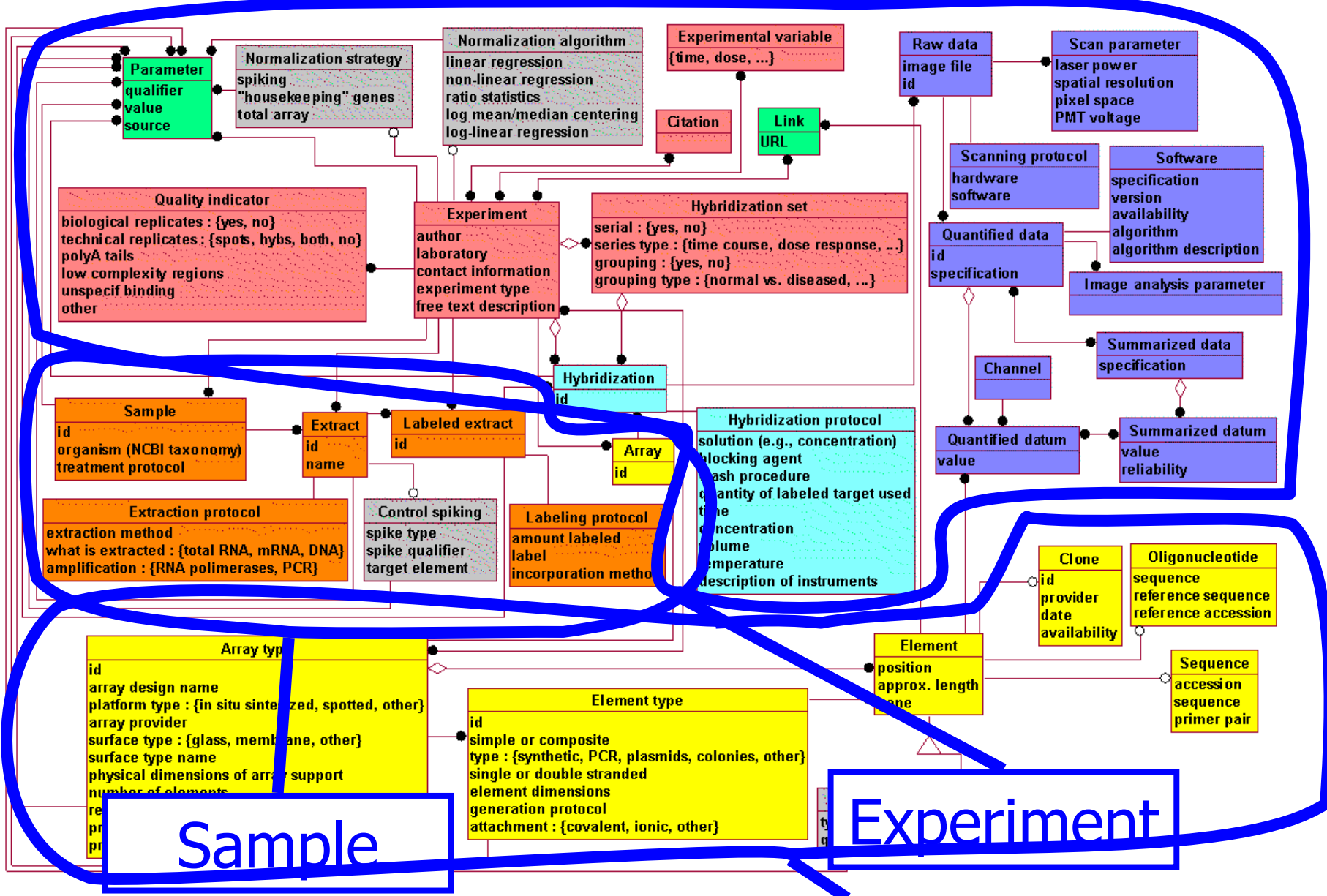
Experimente



MIAME

- MIAME
 - „Minimum Information about a Microarray Experiment“
 - Menge von notwendigen Informationen, um Ergebnisse einzuschätzen und Daten vergleichen zu können
- Sechs Bereiche

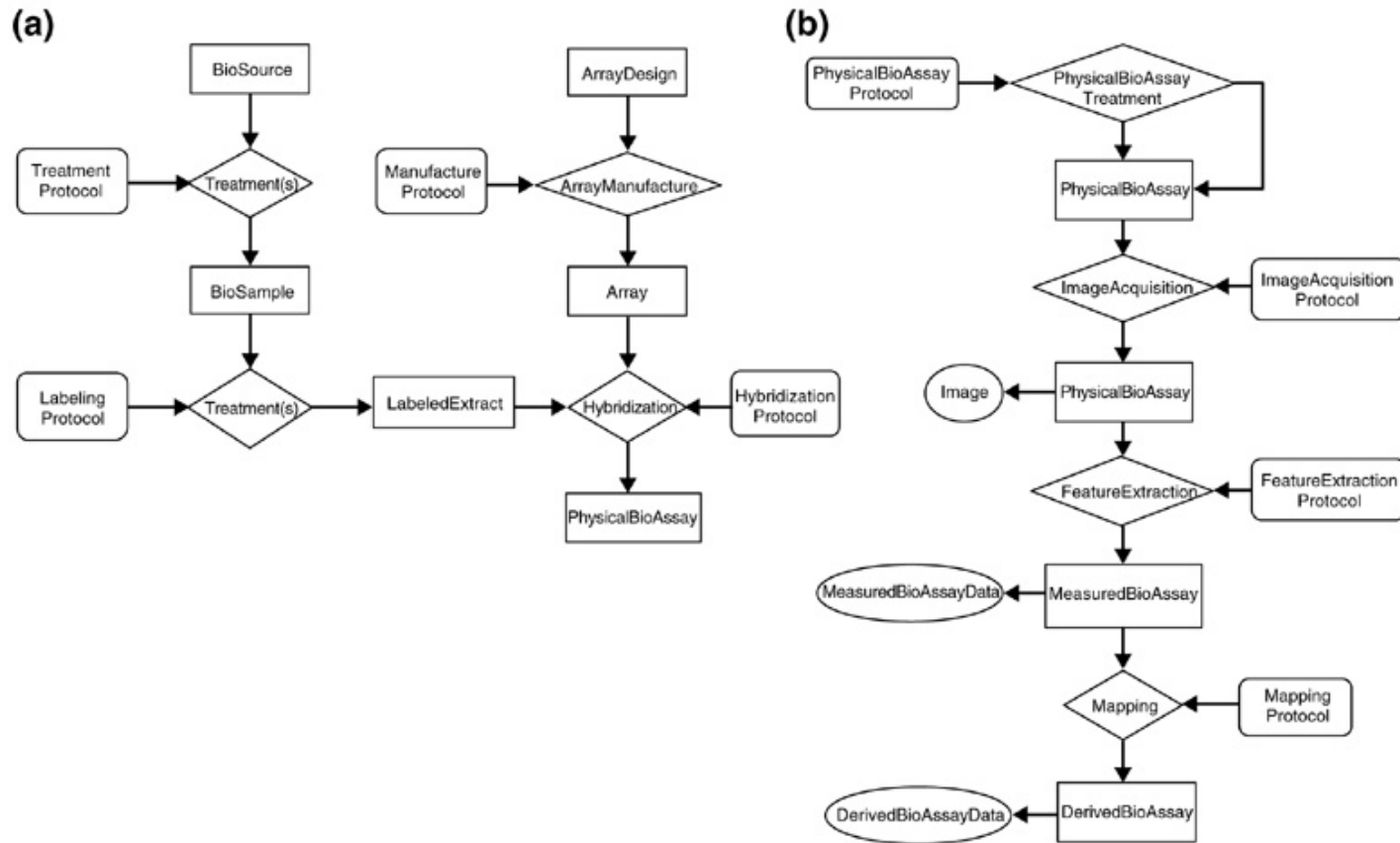
Experimental design	Ziel, Methode, Sampleauswahl, ...
Array Design	Art des Arrays, Layout der Gene, ...
Samples	Taxonomie (Kein Stamm oder Entwicklungsstatus wegen hoher Variabilität zwischen Spezies)
Hybridization	Lösung, Reagenzien, Waschverfahren, ...
Measurement	Bilder und Rohdaten
Normalization	Methode



MAGE

- **MGED**: Microarray Gene Expression Data Society
- **MAGE**: Microarray and Gene Expression
 - Arbeitsgruppe der MGED zu Standards
 - Konstituiert als OMG Working Group
 - Standards
 - MAGE Object Model
 - MAGE Markup Language
 - MAGE OM ist MIAME-compliant
- Implementierung von MAGE-OM: ArrayExpress
- **MAGE-Ontologie**: C.v. für Layouts,, Protokolle, etc.

MAGE Workflow



Quelle: [SMS+02]

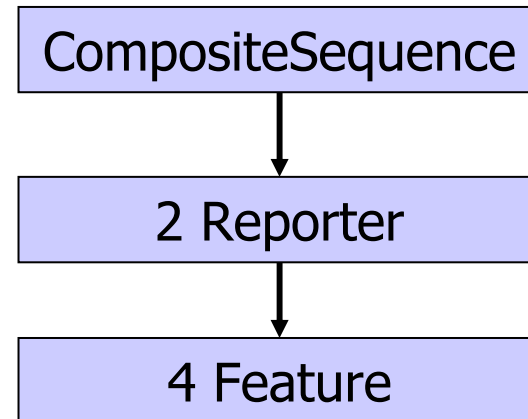
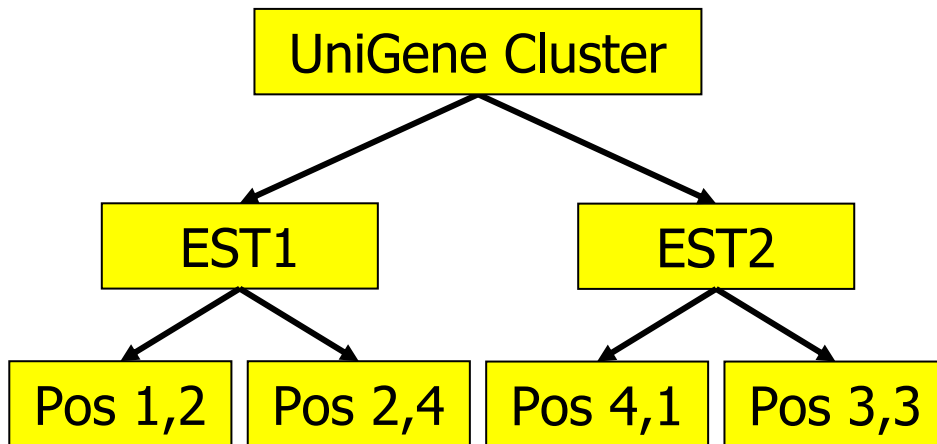
MAGE Objektmodell

132 Klassen, 17 Packages, 150 Seiten Spec.

Experiment	Menge von Hybridisierungen
ArrayDesign	Anordnung von DesignElements
DesignElement	Feature, Reporter, CompSeq
Array	Physikalisch hergestelltes Array
BioAssay	Verarbeitung eines Arrays
Biomaterial	Beschreibung der Sample
Analysis	Auswertung eines Arrays

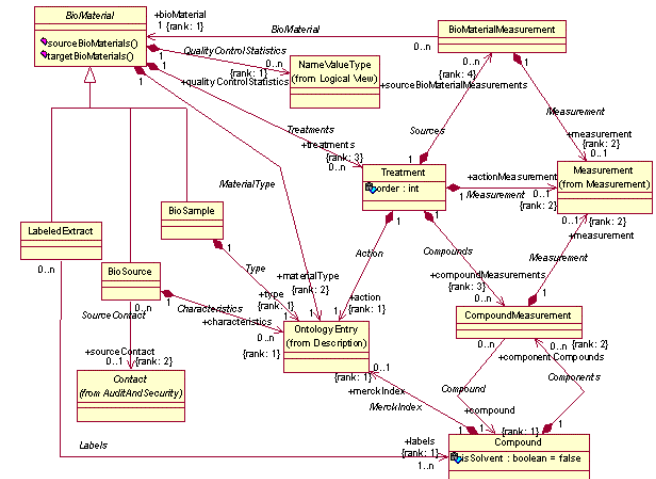
Eigenschaften

- Trennung von Feature, Reporter, CompSeq
 - Feature ist der Spot
 - Reporter ist das physische Objekt
 - CompositeSequence ist Zusammensetzung von Reportern



Weitere Eigenschaften

- ExpressionValueSet
 - Mehrere Werte pro Spot möglich
 - Roh, normalisiert, skaliert, rot/grün, ...
- Beschreibung von Materialien
 - Arraylayout, Typ, Herkunft
 - Sampleherkunft, Herstellungsprozess
- Beschreibung von Protokollen
 - Hybridisierung
 - Proben / Sampleaufbereitung
- (Limitierte) Beschreibung von Analysen
 - Normalisierung, Bilderkennung, Skalierung
 - Programme, Programmversionen, Algorithmen



Bewertung

- MIAME / MAGE: Beginnender Standard für Publikationen
- Sehr aufwändiges Format
 - Manuell kaum handhabbar
 - Sehr große XML Files (Kompression)
 - [Submission Tools, Generierung aus LIMS](#)
- Vorteile
 - Identische Schritte – vergleichbare Daten
 - Wann sind alle Schritte identisch?
 - Wurden Schritte wie beschrieben ausgeführt?
 - [Datenaustausch, externe Validierung, Best Practices](#)
- Nachteile
 - Abweichende Schritte – verloren
 - Erheblicher Overhead für einzelne Labore

Teil IV. Datenbanken

- ArrayExpress
- Genexpression Omnibus
- GeneX (Open Source)

Datenmengen

- Experimente

- ~ 30 000 Gene (homo sapiens)
- ~ 320 Zelltypen (homo sapiens)
- 2000 Compounds (kleiner Screen!)
- 2 Konzentrationen
- 3 Zeitpunkte
- 5 Wiederholungen

- Datenmengen

- ~ 10^{12} Messpunkte \geq 1 TB

Genexpressionsdatenbanken

- **Sehr populär**
 - GeneX/GeneX-Lite, Nat. Cent. for Genome Research
 - maxdSQL, U Manchester
 - Array Express, EBI
 - Express DB, Harvard
 - GeneExpressionBus (GEO), NCBI
 - ArrayDB, Nat. Human Genome Res. Inst.
 - NOMAD, UCSF
 - Stanford Microarray Database (SMD), U Stanford
 - MCHIPS, DKFZ Heidelberg
 - Affymetrix Datenbank
 - ...

Datenbanken

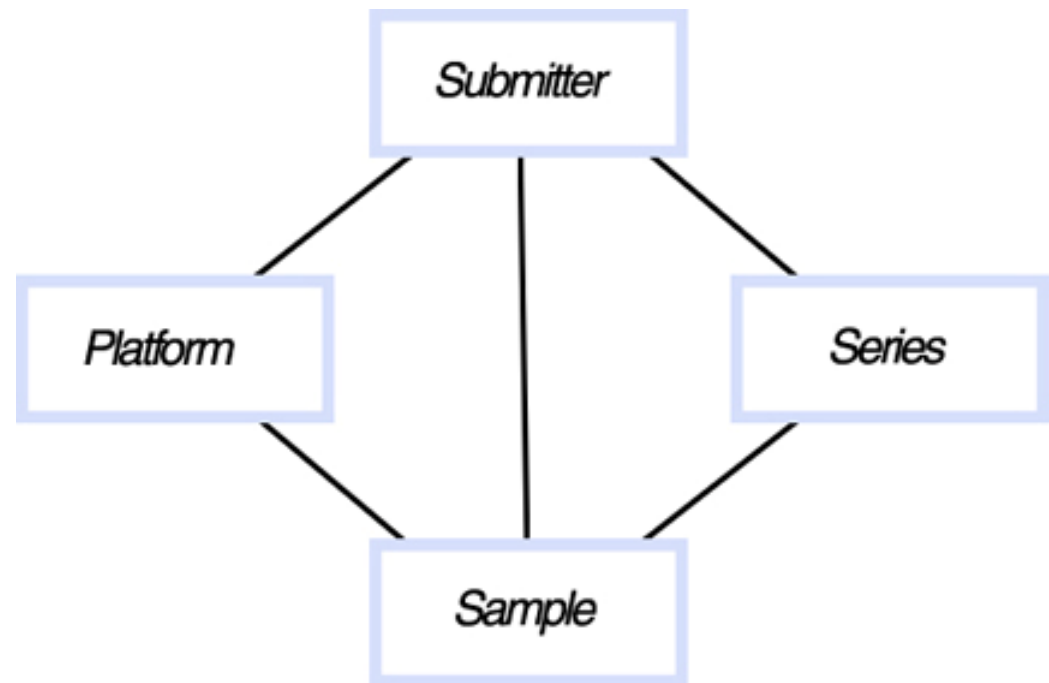
- Kommerziell oder akademisch
- Zur dezentrales Datenrepository oder zur lokalen Installation
- Praktisch alle basierend auf RDBMS (Postgres, Oracle, Sybase)
- Teilweise Integration von Analysesystemen (J-Express, R, Xcluster)
- I.d.R. geringer Datenintegrationsgrad (Xrefs)

Arrayexpress

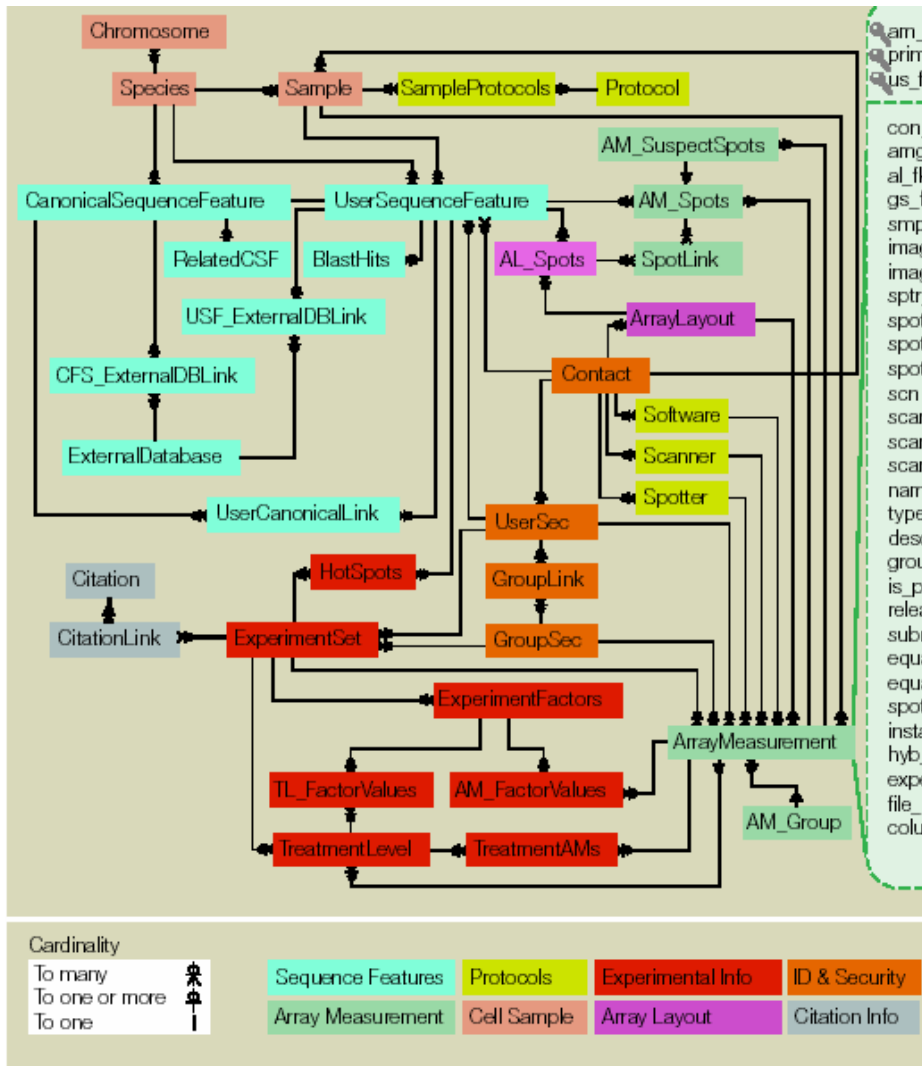
- Öffentliche Datenbank am EBI (Oracle)
- Ziel: Das **EMBL für Genexpressionsdaten**
 - Ähnlich GeneExpressionOmnibus, NCBI
- Bestandteile
 - MAGE / MIAME kompatibel
 - MAGE ML Import und Export
 - Submissiontool
 - Browse & Search (Web)
- Langsam wachsender Datenbestand
 - 4/2003: 16 arrays
 - 7/2994: 213 Experimente, 224 Arraybeschreibungen

Gene Expression Omnibus

- NCBI Public Repository
- RDBMS – Schema verfügbar
- Flatfile Import
- Inhalt (4/2003)
 - 5000 Samples
 - 700 Series
- Keine Oligoarrays



GeneX



- Repository und Open Source Code
 - Linux, Windows
 - Postgres, Oracle
- Ähnlich MAGE
 - Multi-Chanel
 - Reporter – Feature
 - Experimente
 - Controlled Vocs.
- Virtuelle Experimente

Zusammenfassung

- Microarrayexperimente
 - Hohes Potential
 - Schwierige Analyse
- Experimente absichern (teuer!)
 - Experimente mehrmals wiederholen
 - Andere Techniken für Stichpr. verwenden (Blotting, RT-PCR)
- Daten verschiedener Experimente kaum kombinierbar (Methoden, Sample, Arrays, ...)
- Unmenge von Datenbanken verfügbar
- Qualitativer Vergleich kaum bekannt
 - Performancemessungen
 - Toolintegration, OLAP Funktionalität
 - Datenintegration

Literatur

- [BH02] Baldi, P. and Hatfield, W. G. (2002). "DNA Microarray and Gene Expression", Cambridge University Press.
- [DIB97] DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." *Science* **278**: 680-6.
- [DKR03] Do, H. H., Kirsten, T. and Rahm, E. (2003). "Comparative Evaluation of Microarray-based Gene Expression Databases". BTW03..
- [MSZ+01] Mangalam, H., Stewart, J., Zhou, J., et al. (2001). "GeneX: An Open Source Gene Expression Database and Integrated Tool Set." *IBM Systems Journal* **40**(2): 552-569.
- [SD01] Schulze, A. and Downward, J. (2001). "Navigating gene expression using microarray - a technology review." *Nature Cell Biology* **3**.
- [SMS+02] Spellman, P. T., Miller, M., et.al. (2002). "Design and implementation of microarray gene expression markup language (MAGE-ML)." *Genome Biol* **3**(9).
- [BHQ+01] Brazma, A., Hingamp, P., Quackenbush, J., et. al. (2001). "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data." *Nature Genetics* **29**(4): 365-71.
- [BPS03] Brazma, A., Parkinson, H., Sarkans, U., et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." *Nucleic Acids Res* **31**(1): 68-71.