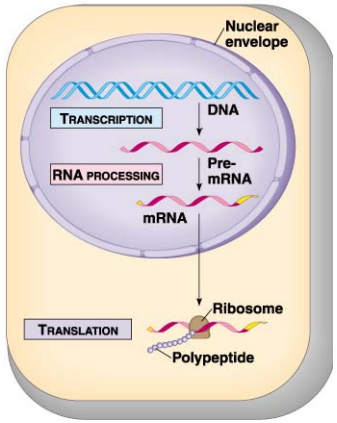
Kapitel 4 (Forts.) Genexpression

- Genexpressionsexperimente
  - Verfahren
  - Anwendungsgebiete
  - Systematische Probleme, Normalisierung
- Analyse von Genexpressionsdaten
  - Differentielle Expression
  - Clustering zur Ko-Expression
- Datenmanagment
  - Übersicht zu bestehenden Genexpressions-DBs
  - Datenarten, Datenmodelle

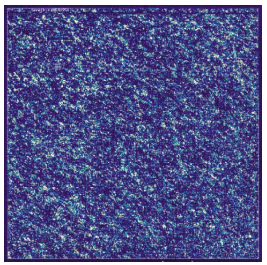
Genexpression

- Was ist Genexpression?
  - Aktivierung der Gentranskription durch endogene, exogene Einflüsse
  - Ausbildung der einem Gen inhärenten Eigenschaften
- Ziele der Genexpressionsanalyse
  - Charakterisierung der Funktion von Genen, deren Interdependenzen, Interaktionen und Einfluss in verschiedenen Netzwerken (metabolische N., regulatorische N. etc.)
- Messung der Genexpression
- Ziele:
  - Messung der RNA Konzentration in Zellen unter verschiedenen Bedingungen (gesundes vs. krankes Gewebe)
  - Suche nach Genen mit gleicher Expression (Koexpression) bzw. differenzieller Expression
- Techniken: Northern Blotting, SAGE, Microarray ...

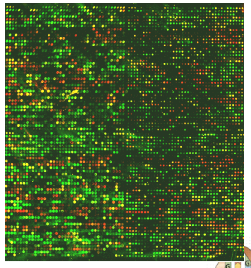


Microarrays

- cDNA Arrays, Oligo Arrays
  - Chiptechnologie (Wafer)
  - 'single stranded' Sequenzen
  - Unterscheidung nach Sequenzart, Sequenzlänge
- Hersteller: Affymetrix, Agilent, Rosetta Inpharmics etc.
- einfarbige vs. zweifarbige Arrays

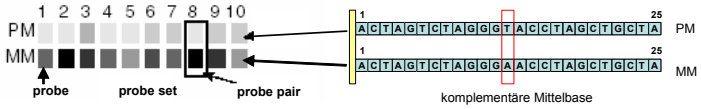


Verteilung von Untersuchungsgewebe + Kontrolle auf einem oder mehreren Chips

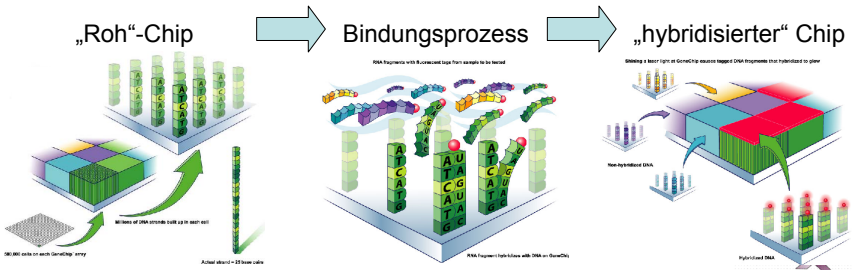


Affymetrix GeneChip Technologie

- verschiedene Chiptypen
  - Abbildung unterschiedlicher Spezies, Transkriptteile
- Terminologie

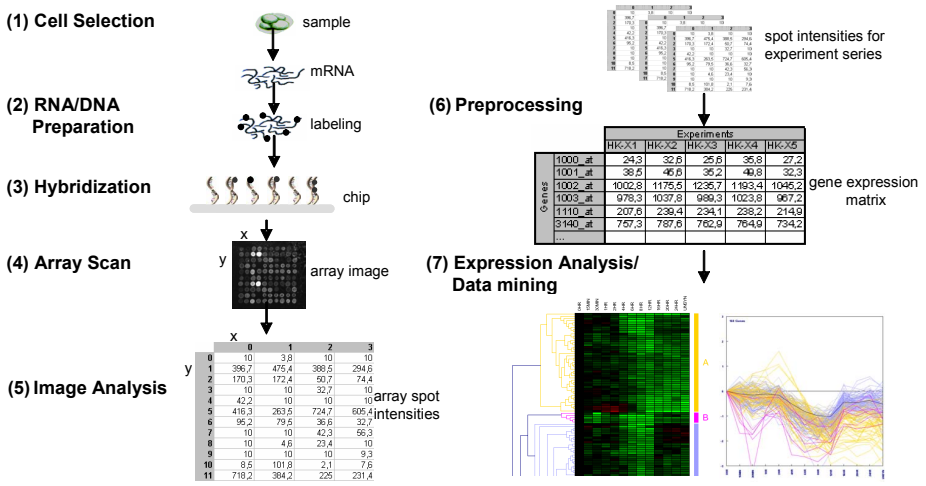


- Hybridisierungsprozess (stark vereinfacht)

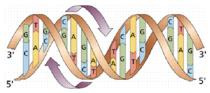


Bildquelle: Affymetrix, http://www.affymetrix.com

Expressionsexperiment und -analyse



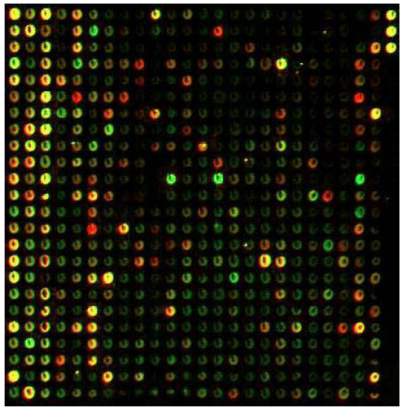
- Analyse ist abhängig vom verfolgten Ziel bzw. der Fragestellung



Zweifarbaufnahmen

- Quantifizierung des Expressionsniveau ist extrem schwierig
  - Warum? Später
- Ziel ist meistens auch nur, Unterschiede in Expression zu finden
  - Ausreichend für Klassifizierung
  - Absolute Werte nicht notwendig
- Zwei Samples auf einem Array
  - Gesund – Krank
  - Unterschiedlicher Farbstoff (rot, grün)
  - Laserabtastung auf zwei Wellenlängen
- Vorteil
  - Unterschiede in Proben, Array, Scanner nivelliert

Ergebnis



- Sample A: rot
- Sample B: grün
- Verhältnis Rot/Grün
  - Dunkel: Gen weder in A noch B exprimiert
  - Rot: Gen nur in A exprimiert
  - Grün: Gen nur in B exprimiert
  - Gelb Gen in A und B exprimiert

Anwendungsgebiete

- Unterschiede in der Genaktivität zwischen Zellen
  - Zwischen Geweben / Zelltypen
    - Nerven, Haut, Muskel, Gehirn, ...
  - Zwischen verschiedenen Spezies
    - Mensch, Maus, Fliege, ...
  - Zwischen verschiedenen Entwicklungsstadien
    - Embryo, Säugling, Jugendlicher, Erwachsener, ...
  - Bei unterschiedlichen Umwelteinflüssen
    - Temperatur, Nahrung, Medikamente, ...
- Ko-Regulation von Genen
  - Gleiche Aktivitätsmuster – gleiche Aufgabe?
  - Gleiche Aktivitätsmuster – gleiche Regulation?

Diagnostik

- Finden typischer Genexpressionsmuster
- Reportergene, Tumormarker
  - Screenen aller Gene, Finden der charakteristischen
- Differentielle Diagnostik von Tumoren
- Personalisierte Medizin
  - Individuelle Medikamentwirksamkeit
  - Pharmokogenomics

➢ Zuordnung von Genexpressionsmustern zu Phänotypen

Systematische Probleme

- Gesund–krank Messung schwierig
  - Genexpression ist in Zellen immer unterschiedlich (Phase in „Cell Cycle", Umgebung, Vorfahren, ...)
  - Unterschiede zwischen zwei gesunden Zellen u.U. größer als zwischen Gesund – Krank
  - Tumore: Schwierig, reine Samples zu bekommen (Tumor–Gesund Gemische)
- Genrepräsentation
  - Viele Gene nur selten und in geringer Dosis aktiv (insbesondere embryonale Zellen)
  - Geringe / fehlende Repräsentation in cDNA Libraries
  - Geringe Menge in Samples – nicht nachweisbares Signal

Systematische Probleme 2

- RNA Isolierung
  - Jede Zellmanipulation zur Verarbeitung induziert Veränderung in Genexpression (Stress, Apoptose, ...)
  - Verfälscht das Ergebnis ungewollt
- Aktivitätszeiträume
  - Signalschritte sind teilweise sehr schnell (<1sec)
  - Wichtige Zwischenschritte in Reaktionsketten werden übersehen
  - Oder: extrem viele Samples notwendig
- Ursache – Wirkung nicht trennbar
  - Primär-, Sekundäreffekte: Tumor (primär) führt zu erhöhter Zellteilung (sekundär) mit 100en aktivierten Genen

➢ Schwierig, charakteristische Effekte zu finden

➢ Schwierig, Vergleichbarkeit von Daten herzustellen

Normalisierung

- Intensitäten verschiedener Experimente sind nicht vergleichbar
  - Anzahl Zellen zur Sampleaufbereitung
  - Menge von mRNA in Zellen
  - Experimentelle Parameter (Temperatur, Chemikalien, Dauer, ...)
  - Sensitivität der Messung (Kamera, Laser)
- Normalisieren auf mRNA Menge in Sample
  - Messen der totalen mRNA Menge in Sample
  - Teilen aller Intensitäten durch diesen Wert
  - Annahme: „Zellen produzieren proportionale RNA Mengen"
- Referenzgene
  - Auswahl von „Housekeeping" Genen
  - Teilen aller Intensitäten durch deren Intensität
  - Annahme: „Bestimmte Gene sind immer gleich exprimiert"

Normalisierung

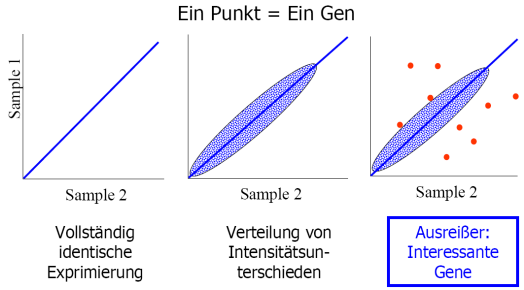
- Referenz RNA
  - Zugabe von festen Mengen bekannter RNA zum Sample
  - Teilen aller Intensitäten durch gemessene Intensitäten dieser Peaks
  - Unterschiede im Protokoll nach der Zugabe können nivelliert werden (zu spät...)
- Globale Skalierung
  - Summe aller Intensitäten in Array berechnen
  - Teilen aller Intensitäten durch diese Summe
  - Reine Skalierung, Proportionen bleiben erhalten, absolute Werte bedeutungslos
  - Gewährleistet bestenfalls Vergleichbarkeit innerhalb eines festen Protokolls
    - Aber das mit wenig Zusatzannahmen

Vergleich Genexpression - Sequenzierung

- Genomsequenzierung
  - Sequenz ist stabil
  - Praktisch identisch innerhalb einer Spezies
  - Einmal sequenziert – für immer richtig
  - Sequenz ist „richtig" – nur kleine Fehler
- Genexpression
  - Abhängig von vielen Faktoren: Zelltyp, Umgebung, Vergangenheit, Entwicklungsstufe, Eltern, ...
  - Messungen schwer vergleichbar, da nie alle Umgebungsvariablen gleich sind
  - Eine „normale" Genexpression gibt es nicht

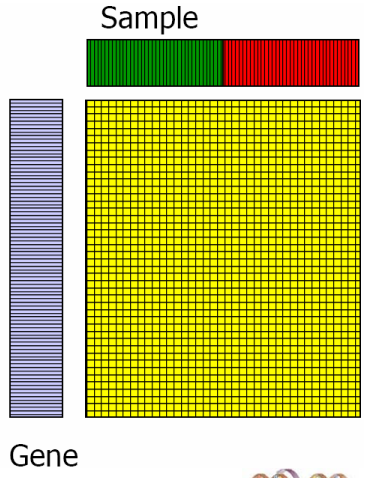
Analyse von Genexpressionsdaten

- Differentielle Expression
- Rohdaten
  - Expressionsintensitäten einzelner Gene
  - Experimentreihen: Sample 1(Kontrolle), Sample 2 (Krank)



Differentielle Exprimierung

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



Simple Fold

– X differentiell exprimiert, gdw.:

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

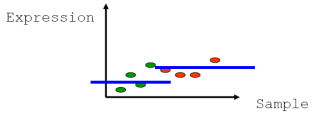
– log: Gleichbehandlung von Steigerung/Verringerung

– Signifikanz wird definiert durch Schwellwerte t, z.B.

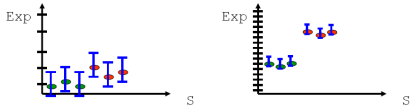
- <2: Uninteressante Veränderung
- 2-4: interessant
- >4: sehr interessant

Probleme des Simple Fold

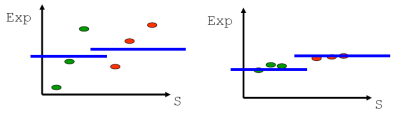
- Vergleicht nur die Mittelwerte



- Unabhängig von absoluten Größen und Fehlerraten



- 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-Tests

- 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 Exprimierung
- X differentiell exprimiert mit Signifikanz α gdw. $|t| > STUDENT(\alpha)$
- STUDENT(α): 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

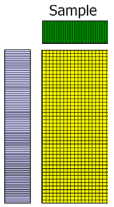
- Signifikanz der Aussage abhängig von gewünschter false-negative Rate
- P-Value: kleinstes α, für das T-Test noch signifikant

Differentielle Exprimierung

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

Ko-Regulation

- Bisher: Erkennen des auffälligen Verhaltens eines Gens
- Jetzt: Erkennen, welche Gene gemeinsam auf einen Stimulus reag.
- Finden von Genen mit „gleicher" Veränderung
  - Welche Gene reagieren gemeinsam auf Temperaturstress?
  - Gruppierung nicht bekannt
  - Clustering – „Unsupervised learning"
  - Gen X = $\{s_{x1},...s_{xn}\}$
  - Gen Y = $\{s_{y1},...s_{yn}\}$
  - ➢ Punkte im n-dimensionalen Raum
- Ähnlichkeitsmaße zwischen zwei Genen
  - Euklidscher Abstand $\sqrt{(s_{x1}-s_{y1})^2+...+(s_{xn}-s_{yn})^2}$
  - Pearson's Korrelations Koeffizient $r=\frac{\sum_{i=1}^{n}(s_{xi}-\bar{s}_x)(s_{yi}-\bar{s}_y)}{\sqrt{\sum_{i=1}^{n}(s_{xi}-\bar{s}_x)^2\sum_{i=1}^{n}(s_{yi}-\bar{s}_y)^2}}$
- ➢ Cluster: Gruppen mit
  - Hoher Ähnlichkeit zwischen Mitgliedern
  - Geringer Ähnlichkeit zu Nicht-Mitgliedern



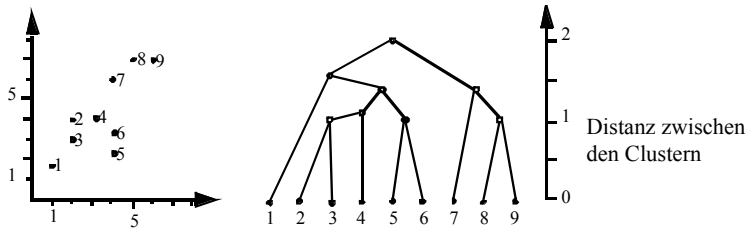
Hierarchische Clusterverfahren

•Ziel
- Konstruktion einer Hierarchie von Clustern (*Dendrogramm*), so daß immer die Cluster mit minimaler Distanz verschmolzen werden

•Dendrogramm
- ein Baum, dessen Knoten jeweils ein Cluster repräsentieren, mit folgenden
- Eigenschaften:
  - die Wurzel repräsentiert alle Gene
  - die Blätter repräsentieren einzelne Gene
  - ein innerer Knoten repräsentiert die Vereinigung aller Gene, die im darunterliegenden Teilbaum repräsentiert werden

Hierarchische Clusterverfahren

•Beispiel eines Dendrogramms



•Typen von hierarchischen Verfahren

• Bottom-Up Konstruktion des Dendrogramms (*agglomerative*)

• Top-Down Konstruktion des Dendrogramms (*divisive*)

Hierarchische Clusterverfahren

Agglomeratives hierarchisches Clustering (bottom up)

1. Bilde initiale Cluster, die jeweils aus einem Gen bestehen, und bestimme die Distanzen zwischen allen Paaren dieser Cluster.
2. Bilde einen neuen Cluster aus den zwei Clustern, welche die geringste Distanz zueinander haben.
3. Bestimme die Distanz zwischen dem neuen Cluster und allen anderen Clustern.
4. Wenn alle Gene sich in einem einzigen Cluster befinden: Fertig, andernfalls wiederhole ab Schritt 2.

Hierarchische Clusterverfahren

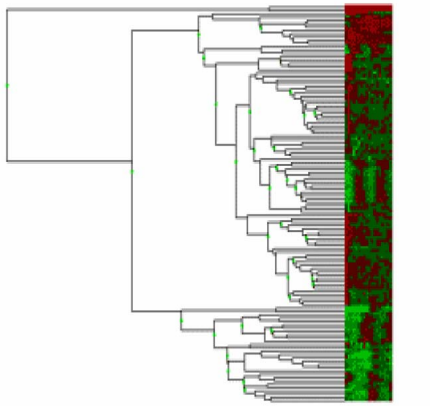
•*Generische Distanzfunktionen für Cluster*

• Sei eine Distanzfunktion *dist*(*x*,*y*) für Paare von Genen

• Seien *X*, *Y* Cluster, d.h. Mengen von Objekten.

- *Centroid-Link* $\text{centroidLinkDist}(X,Y) = \text{dist}(\bar{x},\bar{y}),\ \bar{x} = \frac{1}{|X|}\sum_{x\in X} x,\ \bar{y} = \frac{1}{|Y|}\sum_{y\in Y} y$
- *Single-Link* $\text{singleLinkDist}(X,Y) = \min_{x\in X, y\in Y} dist(x,y)$
- *Complete-Link* $\text{completeLinkDist}(X,Y) = \max_{x\in X, y\in Y} \text{dist}(x,y)$
- *Average-Link* $\text{averageLinkDist}(X,Y) = \frac{1}{|X|\cdot|Y|}\cdot \sum_{x\in X, y\in Y} dist(x,y)$

Reale Daten



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

Komplexität und Bewertung

- 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: $O(n^2)$
- 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)

Fuzzy k-Means [GE02]

- Idee
  - Model für K überlappende Clusters, Zuordnung der Gene durch member-scores
  - Ein Gen kann in mehreren Clustern Mitglied sein
- Minimiere Zielfunktion:

$$J(F,V) = \sum_{i=1}^{N}\sum_{j=1}^{K} m_{XiVj}^2 \, d_{XiVj}^2$$

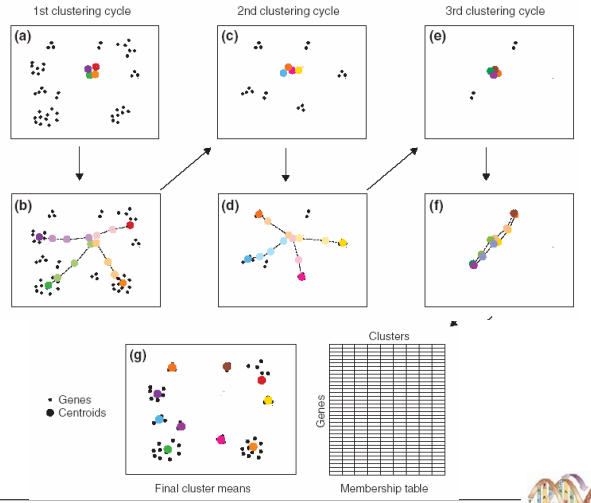
  - X_i … Expressionsmuster des i-ten Gens
  - V_j … Repräsentant von Cluster j
  - d_{XiVj} … Pearson Korrelation (zw. -1 und 1)
  - m_{XiVj} … Membership-Score von Xi in Cluster j

$$m_{XiVj} = \frac{\frac{1}{d_{XiVj}^2}}{\sum_{j=1}^{K}\frac{1}{d_{XiVj}^2}}$$

Fuzzy k Means

- Heuristik um einen guten Wert für k zu bestimmen
- Drei Zyklen
  1. Init. Zentroide auf Eigenvektoren, bestimme Membership und verschiebe Zentroide zum gewichteten Durchschnitt bis das Verfahren konvergiert
  2. Fasse ähnliche Zentroide zusammen (Pearson corr. >0.9), entferne Gene mit Pearson Korr. >0.7, füge neue Zentroide hinzu
  3. Wiederhole Schritt 2
- Abschluß: bestimme Mitgliedschaft der Gene zu den gefundenen Clustern

Fuzzy k Means Beispiel

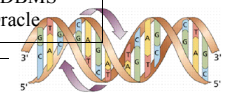


Zusammenfassung

- Datenanalyse
  - Differentielle Expression: Signifikanz von Änderungen
    - Simple Fold
    - T-Test
  - Ko-Regulation von Genen: Gemeinsame Aufgaben
    - Hierarchisches Clustering
    - K-Means
    - Bi-Clustering ...
- Viele Methoden und Implementierungen (Exel, R, ...)

Beispiele bestehender GE-Datenbanken

Name	Organisation / Institut	DBMS
ArrayDB	National Human Genome Research Institute (NHGRI), USA *http://genome.nggri.nih.gov/arraydb*	RDBMS Sybase
ExpressDB	Havard University, USA *http://arep.med.havard.edu/ExpressDB*	RDBMS Sybase
GeneX	National Centre for Genome Resources (NCGR), USA *http://genebox.ncgr.org/genex*	RDBMS Sybase
GIMS	University of Manchester, GB *http://www.cs.man.ac.uk/~norm/gims*	ODBMS Poet
M-CHIPS	German Cancer Research Centre, Germany *http://www.mchips.de*	RDBMS PostGreSQL
RAD2	University of Pennsylvania, USA *http://www.cbil.upenn.edu/RAD2*	RDBMS Oracle
SMD	Stanford University, USA *http://genome-www4.stanford.edu/MicroArray/SMD*	RDBMS Oracle
YMD	Yale University, USA *http://info.med.yale.edu/microarray*	RDBMS Oracle

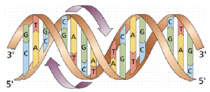


Datenarten

- Verschiedene Arten von Genexpressionsdaten mit unterschiedlicher Charakteristik und Anforderungen erfordern differenzierte Sichtweise

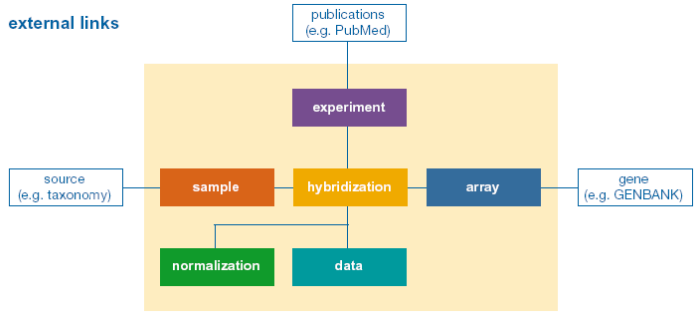
Datenart		Quelle	Datentyp	Charakteristik	Nutzung
Bilddaten		Experiment, Scanvorgang	binär	große Dateien (>20MB)	Generierung von Expressionsdaten
Expressionsdaten		Bildanalyse	ASCII, Zahlenformat	schnell wachsende Menge	statistische Analyse (Clustering), Visualisierung und
Annotationsdaten	Experiment & Sample-annotation	Benutzereingabe	Text	manuelle Eingabe, oft Textfelder	Integration in GE-Analyse, notwendig zur Interpretation der Analyseergebnisse
	Genannotation	externe, öffentliche Quellen		regelmäßge Aktualisierungen in den Datenquellen	

- vielfach keine Speicherung der Bilddaten
- Management von Daten mehrerer Genexpressionstechniken



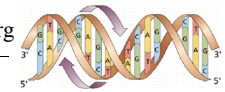
Experimentannotationen

- Dokumentation des experimentellen Prozesses
- vielfach Freitext, keine Benutzung abgestimmter bzw. standardisierter Vokabulare
- "Minimal Information About Microarray Experiment" - MIAME Standard

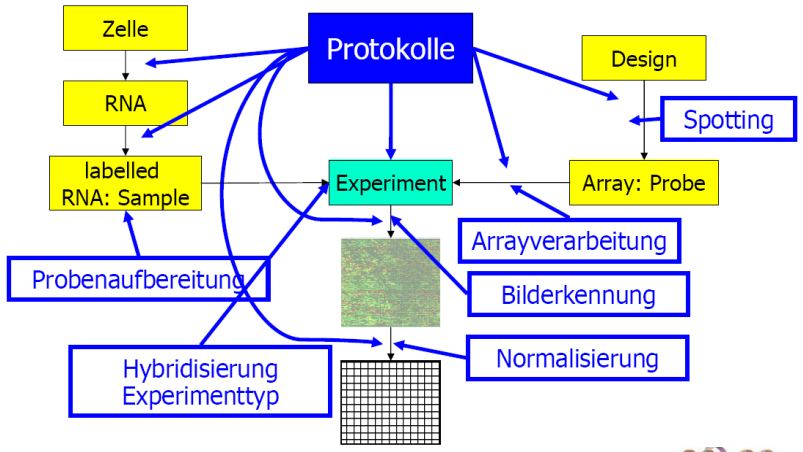


Bildquelle: MGED

  - Umfang für spezielle Domains nicht ausreichend → MIAME/Tox (Toxicogenomics) u.a.
  - Datenaustausch per MAGE-ML (MAGE-OM)
  - "Microarray Gene Expression Data" (MGED) Society → http://www.mged.org



Prozesse



MIAME

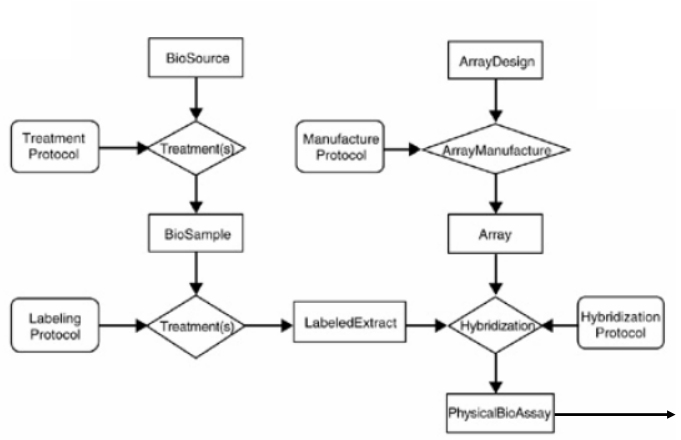
- MIAME
  - „Minimum Information about a Microarray Experiment"
  - Menge von notwendigen Informationen, um Ergebnisse einzuschätzen und Daten vergleichen zu können
- Sechs Bereiche

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

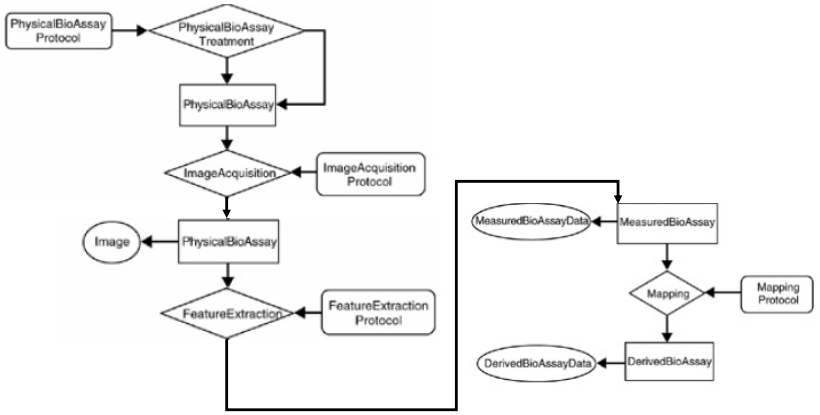
MAGE

- 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



Mage Workflow (2)



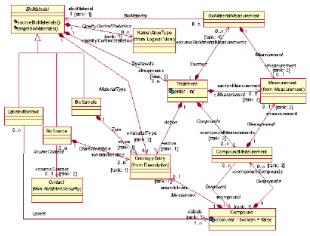
MAGE Objektmodell

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

- 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



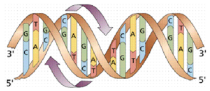
Bewertung

- 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

Mechanismen zur Datenintegration

- Virtuelle Integration: Web Links und föderierte DBS
- Web Links auf Basis spezifischer Identifikatoren
  - weit verbreitete Navigation per Link
  - Beispiel einer URL: http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=75212
- Föderierte Datenbanksysteme
  - Schema Intergation (globales Schema generiert aus lokalen Schemas)
  - On-the-Fly Datenintegration: Transformation, Bereinigung, Herstellen der Relation (Join)
  - kaum Anwendung, aber spezifische Tools wie Discovery Link (IBM)
- Materialisierte Integration (Data Warehouse)
  - lokale/zentralisierte Speicherung aller Expressionsdaten und notwendigen (!) Annotationsdaten
  - Stanford Microarray Database (SMD), sonst kaum Anwendung
- Hybride Ansätze
  - Kombination von materialiserter und föderierter Integration
  - speziefische Systeme: SRS (Lion BioScience), BioMax

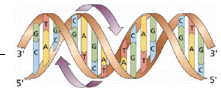
weiterführende Information:
siehe Kapitel Datenintegration



Systemvergleich

	GeneX	M-CHIPS	RAD2	SMD
Datenarten				
Images	nein	nein	nein	Dateisystem
Arrays	cDNA, Oligo, SAGE	cDNA, Oligo, SAGE	cDNA, Oligo, SAGE	cDNA
Experiment Ann.	Geschlecht, Alter, Gewebe, Stadium, ..., Hardware and Softwareparameter	Sehr umfassendes Annotationsschema	Geschlecht, Alter, Krankh., Stadium, ..., RNA Amplifikation, Labeling Protokoll, Scanparameter	Geschlecht, Alter, Status, ...
Vokabulare	lokale Vokabulare	lokale Vokabulare	Standardvokabulare	lokale Vokabulare
Integrationsform				
Web Link	SGD, MGD, dbEST, GenBank, KEGG, SwissProt	GenBank	GenBank, AllGenes, KEGG	dbEST, GeneMap, LocusLink, SwissProt,
föderiert	nein			
Materialisiert	nein	GO functions	nein	GO Funktionen (SGD), Gennamen (WormPD), UniGene
Auto. Update	-	nein	-	ja
Datenanalyse				
Software Tools	RClust, Eisen, CyberT (Web)	proprietär	nein	XCluster (Web)
Integration	Datenbank API	Datenbank API		Datenbank API
Data Mining	Hier., K-means, PCA	Korrespondenzanalyse, Hier. Clustering	nein	Hier., K-means, SOM, SVD
Statistik	T-Tests, Bonferonni Korrektur, …	nein	nein	nein
Visualisierung	interaktive Dendrogramme, Clusterbäume	Korrespondenzanalyse Biplot	-	zoombare Punktgraphiken, interaktive Clusterbilder

Quelle: Do, Kirsten, Rahm, Proc. 10th BTW. 2003



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