Gene Expression Data

Introduction to gene expression data
Expression data storage concept
An example of storage and retrieval: CleanEx
Online Analysis tools for gene expression data
Outline

- Gene expression measurements: from gene-scale to genome-scale
- Data storage: aims, bottlenecks, solutions
- Example of gene expression databases
- Data retrieval systems
- CleanEx: The in-house gene expression database
- Data organization in CleanEx
- Data retrieval in CleanEx
- Examples of online analysis tools
Central Dogma of Molecular Biology

Transcriptome: Genes

Proteome: Proteins

Gene expression measurement
Gene Expression Measurement Methods

Low-Throughput Methods:

Northern Blotting

Typically, measures are done for one gene at a time.
Gene Expression Measurement Methods

**High-Throughput Methods:**

- Whole transcriptome analysis: thousands of genes are studied at the same time
- New problems raised: gene mapping, data cleaning ...
- Need for large-scale pre- and post-processing data analysis
- Need for coherent data management (storage and retrieval systems)
What are high-throughput gene expression measurement methods?

Various technological choices:
• $10^4$ to $10^6$ features on a single array
• Single- vs two-color approach
• Hybridization protocols
• Array or tag sequencing and count

Questions addressed:
• What are the **differences** (in gene expression) between cell lines?
• What is the **difference** between knock-out and wild-type mice?
• What is the **difference** between a tumor and a healthy tissue?
• Are there **different** tumor types?

Key concept:
**Compare** gene expression in two (or more) cell/tissue types?

Gene expression assessed by measuring the number of RNA transcripts in a tissue sample.
RNA abundance in mammalian cells

- rRNA: 80%
- tRNA: 20%
- mRNA: 1%

Molecules/cell:
- 500+ molecules/cell
- 50-500 molecules/cell
- 1-50 molecules/cell

3 x 10^6 molecules/cell
1-2 x 10^4 different genes
Genomics Fundamentals - Complexity

Difficulties:
- Contaminations
- Alternative Splicing
- Alternative PolyAdenylation

mRNA purification
Gene Expression Measurement Methods

**High-Throughput Methods:**

Dual channel arrays
- cDNA microarray
- 60 mer oligoarrays

Single channel arrays
- Affymetrix
- 20 mer oligoarrays

Sequence counts
- Tag counts (SAGE, MPSS)
- EST counts per library
Biological question (e.g. Differentially expressed genes, Sample class prediction, etc.)

Experimental design (chip...)

Microarray experiment

Image analysis
Quality assessment

Normalization

Data Analysis

Estimation  Testing  Clustering  Discrimination

Biological verification and interpretation
Biological question
(e.g. Differentially expressed genes, Sample class prediction, etc.)

Experimental design

SAGE/MPSS experiment

Tags count

Normalization

Data Analysis

Estimation
Testing
Clustering
Discrimination

Biological verification and interpretation
Spotted array preparation

“Average” mouse mRNA

RT-PCR (conversion mRNA-cDNA, amplification)

cDNA isolation

Test sequence (probe) production

~100 - ~2000 bp
Oligo array preparation (e.g. Agilent)

- Sequence databases
- Probe (sequence) design
  - known genes
  - putative genes
  - alternative splicing
  - GC contents
- Gene-specific sequences (~60 bp sequences)
- In-situ synthesis

Millions of experiences worldwide
Affymetrix chip preparation

Sequence databases
- Sequence clusters databases
- GenBank, EMBL, Unigene

Bioinformatics thinking
yields gene-specific sequences (3’-end)

In-situ synthesis
25 nt sequences (probes)

Millions of experiments worldwide

Probe (sequence) design
- known genes
- putative genes
- alternative splicing
- GC contents

11-16 probes = one probeset
~100s of nt “consensus” sequences
High-Throughput Methods : from spot to gene

One spot on array/one tag -> one nucleotide sequence -> one gene ?
High-Throughput Methods: from spot to gene

One spot on array/one tag -> one nucleotide sequence -> one gene?
High-Throughput Methods: from spot to gene

One spot on array/one tag -> one nucleotide sequence -> one gene?

Problems:

★ Regular re-annotation of the sequences spotted on existing chips is needed (cDNA chips, oligochips)

★ One-to-one correspondence between feature and gene is not always correct (All techniques). Difficulties in the numerical data interpretation

★ Alternative splicing might lead to controversial results between two features corresponding to the same gene

★ For Affymetrix chips: All the tags belonging to one probeset might not match the same gene in newer annotations
Gene Expression Measurement Methods

High-Throughput Methods:

- Dual channel arrays
- Single channel arrays
Gene Expression Measurement Methods

High-Throughput Methods:

Tag counts: SAGE

- Extract mRNA from sample
- Transcribe into biotinylated cDNA
- Cut with Anchor Enzyme (AE) and bind to Avidin
- Cut with Tagging Enzyme (TE)
- Ligate and amplify using A & B as primers

Tag counts: MPSS

- Each transcript has a unique signature sequence
- cDNAs with tags are hybridized with microbeads with anti-tags
- Beads with cDNAs are poured and fixed as a single layer in the flow cell
- The signature sequences are compiled from the images obtained at each cycle
- The final matrix contains the tags list and their respective abundance in the sample

Sequence

Process

Data
Global overview

ARRAYS

Array design (gene-to-feature)

Image processing

Normalization

SAGE/MPSS

Sequencing and count

Tag-to-gene mapping

Normalization

Condensation of information

Quality controls at every step

One number per array and per feature/tag

Matrix with one row per feature and one column per sample

To higher level analysis
Dual channel gene expression data

Data on \( p \) genes for \( n \) samples:

<table>
<thead>
<tr>
<th>Genes (Spots)</th>
<th>mRNA samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sample1</td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>-0.45</td>
</tr>
<tr>
<td>5</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

**Gene expression level** of gene \( i \) in mRNA sample \( j \)

\[
= (\text{normalized}) \, \log_2(\text{Red intensity} / \text{Green intensity})
\]
Single channel gene expression data

Data on \( p \) genes for \( n \) samples:

<table>
<thead>
<tr>
<th>Genes (Spots)</th>
<th>mRNA samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>sample1</td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>-0.45</td>
</tr>
<tr>
<td>5</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

**Gene expression level** of gene \( i \) in mRNA sample \( j \)  

\[ \text{Gene expression level} = \underbrace{\text{(normalized)}}_{\text{OR}} \underbrace{\text{Log}_2(\text{Intensity})}_{\text{(normalized)}}(\text{Intensity value}) \]
Counts type gene expression data

Data on $p$ genes for $n$ samples:

<table>
<thead>
<tr>
<th>mRNA samples</th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th>sample4</th>
<th>sample5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>1</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

**Count** of tag $i$ in mRNA sample $j$

= (normalized)(Counts)

**OR**

(normalized)(tag $i$ counts/total counts) in sample $j$
Fundamental Assumptions Made Using Microarray Technology

- That changes in protein concentrations are directly related to corresponding changes in mRNA concentrations
- That alternative splicing of mRNAs has little impact upon protein expression and cellular phenotype
- That mRNA lifetimes / turnovers are unaltered by changes that occur from intended perturbation
- That all mRNAs, regardless of copy number, are captured and extracted with equal efficiency.
- That expression of mRNAs from constitutive (housekeeping) genes are unaffected by perturbing effect
### High-Throughput Methods: Important Questions

**Array design/Tag-to gene attribution:**

One spot on array/one tag -> one nucleotide sequence -> one gene?

**How to deal with old chips -> Reannotation system**

**Mixing numerical data: what can be compared?**

Ratios

Single intensities

Tag counts

--> Different data measurements!

**Data storage**

Ideal format? MIAME compliant? To what extent?

What to keep? From TIFF images to one single value per feature

Dealing with meta-data: sample information, scanner, etc...

Dealing with data retrieval: Fast retrieval of huge data amount...
Other Specific Applications of DNA Microarray Technology

- Gene expression profiling
- Identification of potential drug targets
- Detection of mutations /polymorphisms (SNPs)
- Sequence changes (insertions / deletions)
- Comparative genomic hybridization (CGH)
- Identification of genomes (bacterial, viral)
Timeline of Recent DNA Microarray Developments

1991: Photolithographic printing (Affymetrix)
1994: First cDNA collections are developed at Stanford
1996: Commercialization of arrays (Affymetrix)
1997: Genome-wide expression monitoring in S. cerevisiae (yeast)
2003: Introduction into clinical practices
2004: Whole human genome on one microarray
2006: Genomic tiling arrays
Emergence of gene expression databases

Very heterogeneous data

- Different techniques (SAGE, Dual channel, Affymetrix, MPSS, Solexa ...)
- Different experiments types (time-course, biopsies, cultivated cells, treatments)
- Each experiment raises one point, no attempt to merge data
- No direct links to official gene annotation data
- Very fast increasing amount of data
- People begin to think about comparing different datasets

- Importance of data storage AND retrieval system
- Need for coordination across expression databases
- First “polyvalent” and searchable databases
Gene Expression Data Storage

- A short historical overview about expression data storage
- Accepted format for gene expression databases
- Official gene expression repositories
  - GEO
  - ArrayExpress
  - CIBEX
- Other important gene expression databases
- Specialized databases
- Data retrieval from public gene expression repositories
Gene Expression Databases: Developing Standards

MGED: The Microarray Gene Expression Data society

★ Founded in 1999 by microarray users and producers (Affymetrix, Stanford, EBI

★ Goals:

★ Establishing standards for data quality, storage, management, annotation and exchange at the genomics, transcriptomics, and proteomics levels

★ Facilitating the creation of tools that leverage these standards

★ Promoting the sharing of high quality, well annotated data within the life sciences and biomedical communities.

★ MGED projects:

★ MIAME (Minimum Information About a Microarray Experiment) standard

★ MAGE: MicroArray and Gene Expression MIAME compliant formats, ontology, and integration tools development

★ Others (data transformation and normalization, FISH standards...
MIAME Standards

MIAME describes the Minimum Information that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.

The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)

2. The final processed (normalised) data for the set of hybridisations in the experiment

3. The essential sample annotation including experimental factors and their values

4. The experimental design including sample data relationships

5. Sufficient annotation of the array

6. The essential laboratory and data processing protocols
MAGE : Microarrays and Gene Expression

Goal : to define all the possible terms which are necessary to completely describe microarray experiments, as well as the relationships linking these terms

Tools :

★ MAGE-OM (Object Model)
★ MAGE-ML (Markup Language)
★ MAGE-tab ((Tab format)
MAGE-OM: examples

**BioAssay**

**Package**

**BioAssay**

**Description**

An abstract class which represents both physical and computational groupings of arrays and biomaterials.

**Attributes**

No attributes are defined for this class.

**SUBCLASSES**

- DerivedBioAssay
- MeasuredBioAssay
- PhysicalBioAssay

**FROM**

- BioAssayDatum
- BioAssayDimension
- BioAssayMap
- Experiment
- ExperimentDesign

**SUPERCLASS**

Identifiable

**CLASS**

bbAssay, luaAssays, sourceBioAssays, topLevelBioAssays, bbAssayFactorValues, channels

**TO**

- Channel
- FactorValue
Contact

Description
A contact is either a person or an organization.

Attributes
- address: String
- email: String
- fax: String
- phone: String
- tollFreePhone: String
- URI: String

Package
AuditAndSecurity

FROM
- ArrayDesign
- ArrayManufacture
- Audit
- BioSource
- Database
- Experiment
- Hardware
- Security
- SecurityGroup
- Software

SUPERCLASS
Identifiable

TO
OntologyEntry

SUBCLASSES
- Organization
- Person
Gene Expression Repositories and Databases

Main expression data repositories
  - SMD : the Stanford Microarray Database
  - CGAP and SAGEmap
  - ExpressDB

MGED recommended gene expression repositories
  - GEO
  - ArrayExpress
  - (CIBEX)

Genes oriented databases
  - GeneCards
  - SOURCE

An in-house expression database : CleanEx
The Stanford Microarray Database

- Historical importance (1999)
- The first repository used on an institutional scale
- Supports dual-channel and Affymetrix chips
- Direct pipeline to ArrayExpress, one MIAME compliant repository
- Provides data filtering and analysis
- Provides individual spot history
- Data retrieval is not evident
Official Gene Expression Repositories

**GEO at the NCBI**

- Largest fully public repository for high-throughput molecular abundance data.
- Online resource for gene expression data browsing, query and retrieval.
- Populated with very heterogenous microarray-based experiments (gene expression analysis, genomic DNA arrays, protein arrays, SAGE or even mass spectrometry data.
- Online data submission system via interactive web-based forms.
- Data stored in the GEO SOFT specific format.
- Organized on the basis of three different levels, namely Platforms, Samples, and Series.
GEO (2) : data organization

- **platform (GPL)** : stores the position and corresponding feature of each probe (spot) such as a GenBank accession number, open reading frame (ORF) name and clone identifier.

- **sample (GSM)** : stores the numerical results obtained for a biological sample under one condition.

- **series (GSE)** : a set of samples corresponding to one publication.

- Special file type : **datasets (GDS)**. Curated series, with pre-calculated data analysis.
GEO (3) : data retrieval

- Series, Samples or Platforms Data download in SOFT format
- Numerical values from series can be retrieved as a tab-delimited matrix
- Datasets selection via the NCBI Entrez data retrieval system (keywords based)
- From Entrez, “profiles” gene-centric data retrieval. The profiles output represents a histogram of expression measurements for one gene across each sample in a single GEO dataset.
Official Gene Expression Repositories

**ArrayExpress at the EBI**

- second largest repository for high-throughput molecular abundance data.

- does not accept SAGE data

- Online data submission system via MIAMExpress submission form, heavy, but strictly MIAME based

- Dedicated pipeline for the Stanford Microarray Data

- Data stored in a strict MIAME format.

- Organized on the basis of three different levels, namely Array, Experiment, and Protocol (=~ Platform, Sample and Serie in GEO)

- Data retrieval: Bulk, datasets retrieval via keywords, and gene-based expression profiles retrieval
Growth of Official Gene Expression Data Repositories

By Experiment

By Hybridization
Genes-Oriented Databases

GOAL : giving access to any available expression measurement corresponding to one gene under one single identifier.

Examples of such databases :

- GeneCards
- SOURCE
- CleanEx
Genes-Oriented Databases : GeneCards

- Contains human genes
- Includes automatically-mined genomic, proteomic and transcriptomic information
- Includes orthologies, disease relationships, SNPs, gene expression, gene function...
- Expression data showed :
  
  GeneNote results (Affymetrix-based experiment on normal human tissues)

  Data from Genatlas (from GNF) on human normal tissues

  SAGE data

  Electronic Northern (ESTs counts per tissue category)
GeneCards: example of result with gene KLK3

Microarray Expression for KLK3
GeneCards: example of result with gene KLK3

Electronic Northern Expression for KLK3

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Clones per gene</th>
<th>Total clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMR</td>
<td>0</td>
<td>14,321</td>
</tr>
<tr>
<td>SPL</td>
<td>0</td>
<td>46,604</td>
</tr>
<tr>
<td>TMS</td>
<td>0</td>
<td>74,361</td>
</tr>
<tr>
<td>BRN</td>
<td>6</td>
<td>436,227</td>
</tr>
<tr>
<td>SPC</td>
<td>0</td>
<td>1,946</td>
</tr>
<tr>
<td>HRT</td>
<td>0</td>
<td>34,933</td>
</tr>
<tr>
<td>MSL</td>
<td>10</td>
<td>53,585</td>
</tr>
<tr>
<td>LVR</td>
<td>0</td>
<td>99,980</td>
</tr>
<tr>
<td>PNC</td>
<td>0</td>
<td>69,827</td>
</tr>
<tr>
<td>PST</td>
<td>920</td>
<td>107,672</td>
</tr>
<tr>
<td>KDN</td>
<td>0</td>
<td>118,847</td>
</tr>
<tr>
<td>LNG</td>
<td>0</td>
<td>121,380</td>
</tr>
</tbody>
</table>
GeneCards: example of result with gene KLK3

SAGE Expression for the Best Matching Tag: GGATGGGGGAT

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Tags per gene</th>
<th>Total tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMR Bone marrow</td>
<td>0</td>
<td>36,577</td>
</tr>
<tr>
<td>spl* Spleen</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tms* Thymus</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BRN Brain</td>
<td>1</td>
<td>427,603</td>
</tr>
<tr>
<td>SPC Spinal cord</td>
<td>1</td>
<td>54,785</td>
</tr>
<tr>
<td>HRT Heart</td>
<td>0</td>
<td>63,063</td>
</tr>
<tr>
<td>MSL Skeletal muscle</td>
<td>0</td>
<td>107,836</td>
</tr>
<tr>
<td>LVR Liver</td>
<td>0</td>
<td>66,308</td>
</tr>
<tr>
<td>PNC Pancreas</td>
<td>0</td>
<td>43,040</td>
</tr>
<tr>
<td>PST Prostate</td>
<td>210</td>
<td>123,335</td>
</tr>
<tr>
<td>KDN Kidney</td>
<td>0</td>
<td>40,993</td>
</tr>
<tr>
<td>LNG Lung</td>
<td>0</td>
<td>86,708</td>
</tr>
</tbody>
</table>

* - No libraries exists for that tissue
GeneCards: example of result with gene KLK3

Microarray Integrated Expression by GeneNote and GNF GeneAtlas Data for KLK3
Genes-Oriented Databases: Source

- Based at Stanford, first implemented for link SMD data to genomic information
- Contains human, mouse and rat genes
- Includes clones information for all genes
- Includes an extraction tool for upstream genomic region
- Expression data showed:
  
  Mainly data from the Stanford Microarray database
  
  Expression data from the TissueAtlas (expression in normal tissues)
SOURCE: example of result with gene KLK3

KLK3
Kallikrein-related peptidase 3

UniGene, LocusLink, OMIM, GenAtlas, GeneCard, Ensembl,MapView, AceView, Genome Browser

Aliases

- 3.4.21.77; APS; KLK2A1; PSA; gamma-semionoprotein; hK3; semenogelase; seminin
- ANTIGEN, PROSTATE-SPECIFIC
- P-30 antigen
- PROSTATE-SPECIFIC ANTIGEN
- kallikrein 3, (prostate specific antigen)
- prostate specific antigen isoform 1 preproprotein
- prostate specific antigen isoform 2

Chromosomal Location

Chromosome/Cyto band: 19q13.41

Microarray Gene Expression Data

Data available

Show Gene Expression Data
### 3 Datasets contain expression data for Hs.171995

<table>
<thead>
<tr>
<th>View Expression</th>
<th>Dataset Description</th>
<th>Outside Links</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AML_ALL</strong></td>
<td>Survey of gene expression in Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL). (Golub et al. 1999 <em>Science</em> <strong>286</strong>:531-7).</td>
<td>Authors' Website</td>
</tr>
<tr>
<td><strong>Carcinoma_Classification</strong></td>
<td>Survey of gene expression in a panel of 174 human epithelial tumors, including both primary and metastatic squamous and adenocarcinomas of the prostate, bladder/ureter, breast, colon, gastroesophasus, kidney, liver, ovary, pancreas, and lung. (Su et al. 2001 <em>Cancer Res.</em> <strong>15</strong>:7388-93).</td>
<td>Authors' Website</td>
</tr>
<tr>
<td><strong>NormalTissueAtlas</strong></td>
<td>Survey of gene expression profiles in a diverse group of primary human tissues and organs, as well as transformed cell lines. A total of 101 unique specimens representing 47 tissues/cell lines are represented. (Su et al. 2002 <em>PNAS</em> <strong>99</strong>:4465-70).</td>
<td>Authors' Website</td>
</tr>
</tbody>
</table>
SOURCE : example of result with gene KLK3