Astrid Lægreid

Department of Cancer Research and Molecular Medicine
Norwegian University of Science and Technology

Microarray Core Facility
Norwegian Microarray Consortium
**outline**

- functional genomics
- gene expression
- predicting gene function
- challenges
functional genomics

biological system

- generation of hypotheses
  - biological roles of genes and proteins
- screening genome-wide
- modeling computational biology
- experimental analysis
  - functions of genes and proteins
genome sequencing

human genome
- $3 \times 10^9$ basepairs
- $\sim 35,000$ genes
- $> 100,000$ splice variants

genome-wide screening

- **how?** high-throughput - HTP
- **what?** gene expression, gene-dosage, gene-variation (SNP), protein
- **with?** microarray, mass spectrometry, 2D-gel electrophoresis
functional genomics

biological system

generation of hypotheses
biological roles of genes and proteins

experimental analysis
functions of genes and proteins

screening genome-wide

modeling computational biology
chromosome 21

- 127 known genes
- 98 unknown genes

50 million bases
functional genomics

biological system

generation of hypoteses
biological roles of genes and proteins

experimental analysis
functions of genes and proteins

screening genome-wide

modeling computational biology
from genome to organism

gene

gene copy (mRNA)

protein

organism
from genome to organism

gene

gene copy (mRNA)

protein

cell

organs and tissues

organism
from genome to organism

gene expression

organism

genome
all cell types contain the same genome
..but differ in gene expression patterns....

gene expression determines what you are....
to live is to interact with the environment

challenge

environment

organism

response
the organism responds by making new proteins

or by stop making some of the old ones.....
gene expression depends on

cell type and cell state

we can learn more by measuring

expression of all genes
by measuring changes in gene expression we can discover genes participating in a given biological response
measure thousands of genes and proteins in high throughput analyses

mRNA profiling  protein profiling

microarray  2D gel electrophoresis
mass spectrometry
why measure mRNA?

because DNA microarray is the most high throughput method that can measure gene expression with high sensitivity and specificity
functional genomics

biological system

- generation of hypotheses
  - biological roles of genes and proteins
- screening genome-wide
- experimental analysis
  - functions of genes and proteins
- modeling
  - computational biology
DNA microarray

DNA molecules = specific probe

microscopic slide

5,000 - 80,000 probes pr. array
microarray formats

- cDNA (500-1500 bp)
- long oligonucleotides (40-70-mers)
- short oligonucleotides (20-25-mers)
microarray analysis

1. sample / control
   - RNA isolation
   - labeling

2. hybridization

3. scanning
   - laser 1
   - laser 2
   - red = "up"
   - green = "down"
microarray analysis

Bowtell, Nature Genetics, Supplement, 21:25, 1999
microarray analysis

Bowtell, Nature Genetics, Supplement, 21:25, 1999
functional genomics

biological system

generation of hypotheses
biological roles of genes and proteins

screening
genome-wide

experimental analysis
functions of genes and proteins

modeling
computational biology
biological system

- generation of hypotheses
- experimental analysis
- screening
- modelling
- biological background information
biological system

- generation of hypotheses
- screening
- experimental analysis
- modelling

- gene expression
- gene function
- biological background information
functional classification of genes from time profiles

Astrid Lægreid\textsuperscript{1} and Jan Komorowski\textsuperscript{2}

\textsuperscript{1}Department of Cancer Research and Molecular Medicine
Norwegian University of Science and Technology

\textsuperscript{2}The Linnaeus Centre for Bioinformatics,
Uppsala
The Transcriptional Program in the Response of Human Fibroblasts to Serum


8 hours serum treatment

1, protein disulfide isomerase-related protein
2, IL-8 precursor
3, EST AA057170
4, vascular endothelial growth factor
fibroblast - 24 h serum response

samples for microarray analysis

quiescent non-proliferating

proliferating
dynamic processes

Immediate early
Intermediate
Late

Primary secondary tertiary

Quiescent non-proliferating
Proliferating
molecular mechanisms of transcriptional response

serum = signal

immediate early response factors

immediate early response genes

delayed immediate early response genes

delayed immediate early response genes

secondary transcription factors

intermediate/late response genes

effectors = cellular response
fos - immediate early transcription factor

Transpath; biobase.de
fos - immediate early transcription factor

upstream factors

Transpath; biobase.de
fos - immediate early transcription factor

upstream factors

Transpath; biobase.de
**fos** - immediate early transcription factor

upstream factors

+ downstream genes

Transpath; biobase.de
co-regulation of genes
coding for proteins in a network
in fibroblast serum-response

pro-endothelin → active endothelin → inactive endothelin
co-regulation of genes
coding for proteins in a network
in fibroblast serum-response

+ furin
+ pro-endothelin
+ active endothelin
- CALLA/CD10
- inactive endothelin
cellular processes

- stress response
- protein synthesis
- transcription
- lipid synthesis
- organelle biogenesis
- cell motility
- re-entry cell cycle
- cell re-entry

Quiescent
non-proliferating

Proliferating
fibroblast serum-response transcriptional program

517 gen-probes differential gene expression

497 unique genes

284 known genes

213 unknown genes
Iyer's analysis of transcriptional fibroblast serum response

Expression clusters

Functional clusters
our aim

find relationship between

gene function - gene expression profile
selected challenges in gene-expression analysis

- function similarity corresponds to expression similarity but:
  - functionally correlated genes may be expression-wise dissimilar (e.g. anti-coregulated)
  - genes usually have multiple function
  - measurements may be approximate and contradictory

- can we obtain clusters of biologically related genes?

- can we build models that classify unknown genes to functional classes, that are human legible, and that handle approximate and often contradictory data?

- how can we re-use biological knowledge?
methodology

1. Mining functional classes from an ontology

2. Extracting features for learning

3. Inducing minimal decision rules using rough sets

0 - 4 (Increasing) AND 6 - 10 (Decreasing) AND 14 - 18 (Constant) => GO (cell proliferation)

4. The function of unknown genes is predicted using the rules

Gene Ontology

* The homepage of Ashburner's Gene Ontology: http://genome-www.standford.edu/GO/
## Annotation of Known Genes

<table>
<thead>
<tr>
<th>GENE SYMBOL</th>
<th>GENE NAME</th>
<th>GENEBANK ACCESSION NUMBER</th>
<th>ANNOTATIONS AT THE MOST SPECIFIC LEVEL OF GO</th>
<th>ANNOTATIONS TO THE 23 BROAD CELLULAR PROCESSES USED FOR LEARNING</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEPP1</td>
<td>selenoprotein P, plasma, 1</td>
<td>AA045003</td>
<td>oxidative stress response(GO:0006979), metal ion transport(GO:0006823)</td>
<td>stress response(GO:0006950), transport(GO:0006810)</td>
</tr>
<tr>
<td>EPB41L2</td>
<td>erythrocyte membrane protein band 4.1-like 2</td>
<td>W88572</td>
<td>positive control of cell proliferation(GO:0008284)</td>
<td>cell proliferation(GO:0008283)</td>
</tr>
<tr>
<td>OA48-18</td>
<td>acid-inducible phosphoprotein</td>
<td>AA029909</td>
<td>cell proliferation(GO:0008283)</td>
<td>cell proliferation(GO:0008283)</td>
</tr>
<tr>
<td>CTSK</td>
<td>cathepsin K (pycnodysostosis)</td>
<td>AA044619</td>
<td>proteolysis and peptidolysis(GO:0006508)</td>
<td>protein metabolism and modification(GO:0006411)</td>
</tr>
<tr>
<td>CPT1B</td>
<td>carnitine palmitoyltransferase I, muscle</td>
<td>W89012</td>
<td>fatty acid beta-oxidation(GO:0006635)</td>
<td>lipid metabolism(GO:0006629)</td>
</tr>
<tr>
<td>CLDN11</td>
<td>claudin 11 (oligodendrocyte transmembrane protein)</td>
<td>N22392</td>
<td>cell adhesion(GO:0007155), substrate-bound cell migration(GO:0006929), cell proliferation(GO:0008283), developmental processes(GO:0007275)</td>
<td>cell adhesion(GO:0007155), cell motility(GO:0006928), cell proliferation(GO:0008283), developmental processes(GO:0007275)</td>
</tr>
<tr>
<td>RPL5</td>
<td>ribosomal protein L5</td>
<td>AA027277</td>
<td>protein biosynthesis(GO:0006412), ribosomal large subunit assembly and maintenance(GO:0000027)</td>
<td>protein metabolism and modification(GO:0006411), cell organization and biogenesis(GO:0006986)</td>
</tr>
<tr>
<td>Homo sapiens clone 23785 mRNA sequence</td>
<td>N32247</td>
<td>calcium-independent cell-cell matrix adhesion(GO:0007161)</td>
<td>cell adhesion(GO:0007155)</td>
<td></td>
</tr>
</tbody>
</table>
time profiles of selected processes
Gene Ontology vs. clusters
template-based feature synthesis

All possible subintervals in the time series

Templates: Increasing, Decreasing, Constant

Gene expression time series data

Groups containing genes matching the same templates over the same subinterval

MATCH

12 measurement points, 55 possible intervals of length >2
cross validation estimates

<table>
<thead>
<tr>
<th>PROCESS</th>
<th>AUC</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion homeostasis</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Protein targeting</td>
<td>0.99</td>
<td>0.03</td>
</tr>
<tr>
<td>Blood coagulation</td>
<td>0.96</td>
<td>0.08</td>
</tr>
<tr>
<td>DNA metabolism</td>
<td>0.94</td>
<td>0.09</td>
</tr>
<tr>
<td>Intracellular signaling cascade</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>Energy pathways</td>
<td>0.93</td>
<td>0.12</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>0.93</td>
<td>0.04</td>
</tr>
<tr>
<td>Oncogenesis</td>
<td>0.92</td>
<td>0.11</td>
</tr>
<tr>
<td>Circulation</td>
<td>0.91</td>
<td>0.11</td>
</tr>
<tr>
<td>Cell death</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>Developmental processes</td>
<td>0.90</td>
<td>0.07</td>
</tr>
<tr>
<td>Transcription</td>
<td>0.88</td>
<td>0.11</td>
</tr>
<tr>
<td>Defense (immune) response</td>
<td>0.88</td>
<td>0.05</td>
</tr>
<tr>
<td>Cell adhesion</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Stress response</td>
<td>0.86</td>
<td>0.15</td>
</tr>
<tr>
<td>Protein metabolism and modification</td>
<td>0.85</td>
<td>0.10</td>
</tr>
<tr>
<td>Cell motility</td>
<td>0.84</td>
<td>0.11</td>
</tr>
<tr>
<td>Cell surface rec linked signal transd</td>
<td>0.82</td>
<td>0.15</td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td>0.81</td>
<td>0.14</td>
</tr>
<tr>
<td>Transport</td>
<td>0.79</td>
<td>0.17</td>
</tr>
<tr>
<td>Cell organization and biogenesis</td>
<td>0.79</td>
<td>0.11</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>0.79</td>
<td>0.06</td>
</tr>
<tr>
<td>Amino acid and derivative metabolism</td>
<td>0.69</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.09</strong></td>
</tr>
</tbody>
</table>

A: Coverage: 84%  
Precision: 50%

B: Coverage: 71%  
Precision: 60%

C: Coverage: 39%  
Precision: 90%

Coverage = TP/(TP+FN)  
Precision = TP/(TP+FP)
# the model

## Annotations, Rules and Classifications

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotated genes</strong></td>
<td>273</td>
</tr>
<tr>
<td>within the 23 broad classes of GO biological process</td>
<td></td>
</tr>
<tr>
<td><strong>Gene probes</strong></td>
<td>284</td>
</tr>
<tr>
<td>associated with the 273 genes within the 23 broad biological process classes</td>
<td></td>
</tr>
<tr>
<td><strong>Training examples</strong></td>
<td>549</td>
</tr>
<tr>
<td>annotations associated with the genes in the 23 broad biological process classes</td>
<td></td>
</tr>
<tr>
<td>co-annotations associated with the genes in the 23 broad biological process classes</td>
<td>444</td>
</tr>
<tr>
<td><strong>Rules</strong> generated from the training examples</td>
<td>18064</td>
</tr>
<tr>
<td><strong>Estimated quality of classifications of unknown genes</strong> (cross-validation estimates)</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>84%</td>
</tr>
<tr>
<td>Specificity</td>
<td>91%</td>
</tr>
<tr>
<td>Fraction of classifications that are correct</td>
<td>49%</td>
</tr>
<tr>
<td><strong>Classifications for unknown (uncharacterized) genes</strong></td>
<td>548</td>
</tr>
<tr>
<td>classifications were obtained for 211 of the 213 unknown genes</td>
<td></td>
</tr>
<tr>
<td><strong>(Re-)Classifications for training examples</strong></td>
<td>728</td>
</tr>
<tr>
<td>True positive classifications</td>
<td>519</td>
</tr>
<tr>
<td>True positive co-classifications</td>
<td>356</td>
</tr>
<tr>
<td>False positive classifications</td>
<td>219</td>
</tr>
<tr>
<td>False negative (missing) classifications</td>
<td>30</td>
</tr>
</tbody>
</table>

For 272 of the 273 training examples at least one correct (re-)classification was obtained
conclusions

- our methodology
  - incorporates background biological knowledge
  - handles well the noise and incompleteness in the microarray data
  - can be objectively evaluated
  - predicts multiple functions per gene
  - can re-classify known genes and provide possible new functions of the known genes
  - can provide hypotheses about the function of unknown genes

- experimental work needs to be done to confirm our predictions

Lægreid A, Hvidsten T, Midelfart H, Komorowski J, Sandvik AK.
Predicting Gene Ontology Biological Process from Temporal Gene Expression Patterns.

Hvidsten TR, Lægreid A, Komorowski J.
Learning rule-based models from gene expression time profiles annotated using Gene Ontology.
*Bioinformatics*, 19:1116-23, 2003
Genomic ROSETTA:
http://www.idi.ntnu.no/~aleks/rosetta
how to improve models for prediction of biological roles of genes/proteins?

- improved computational methods
- more training examples
  - more genes/proteins
  - more measurements per gene/protein (time points, cell types, tissues, states,...)
  - more annotations
    (GO, sequence, protein structure, cell biology, physiology, pathology,...)
many levels of information

Bowtell, Nature Genetics, Supplement, 21:25, 1999
gene copy (mRNA)
protein
gene

~35,000 genes

> 100,000 gene (splice) products

> 100,000 proteins

> 200,000 protein states

each cell expresses
  5.- 15,000 genes
  40.-60,000 proteins

several hundred cell types

many different states per cell

tissues and organs are composed of many different cell types
molecular networks within cells

genome → gene → gene copy (mRNA) → protein → cell → organs and tissues → organism
molecular networks within cells
different cell types interact within organs and tissues
different cell types interact during gastric acid secretion

stomach mucosa

stimuli, (food,...)
interconnection within organism
hormones regulate interactions between organs and tissues
expression profiling in biology

determine molecular mechanisms underlying
  ▶ cell function related to cell type and state
  ▶ physiological functions of organisms
expression profiling in disease management

- discover disease subtypes
- improve disease diagnostics
- improve prognostics/choice of treatment
- discover new drug targets
our focus:

Molecular Mechanisms of the Normal and Diseased Gastrointestinal System

- gastric acid secretion
- hypergastrinemia
- gastric cancer

Arne Sandvik and Astrid Lægreid

Department of Cancer Research and Molecular Medicine, NTNU
gastrointestinal physiology and pathophysiology

- gastric acid secretion

- molecular mechanisms?
- regulators, effectors?
gastrointestinal physiology and pathophysiology

- hypergastrinemia

↑ gastrin
↑ proliferation gastric mucosa
↑ ECL-cells
↑ cancer

- molecular mechanisms?
- regulators, effectors?
gastrointestinal physiology and pathophysiology

- gastric cancer

- classification & prediction
  - subtype diagnostics
  - prognostics
  - optimal treatment
  - early diagnostics

- molecular mechanisms?
- regulators, effectors?
functional genomics

biological system

- generation of hypotheses
  - biological roles of genes and proteins

- experimental analysis
  - functions of genes and proteins

- screening genome-wide

- modeling computational biology
screening

biological system

biological background information

modelling

generation of hypotheses

experimenal analysis
challenges....

Information Bases/Derived-Data Databases
Experimental/Clinical Data
challenges....

Information Bases/Derived-Data Databases

Experimental/Clinical Data

link information from various sources in a relevant way
relational database & tools at NTNU

Input
- GenBank Acc. Nr.
- UniGene Cluster ID`s
- Clone ID`s

Database
- Local Gene Annotation Database
  - Gene Ontology
  - LocusLink
  - UniGene
  - SwissProt
  - Homolo-Gene

Application
- eGOn
- NMC Annotation Database

Output
- Editable GO tree
- Statistical tests
- File export
- Gene Annotations
- File export
challenges....

Information Bases/Derived-Data Databases
Experimental/Clinical Data

mine information from unstructured information sources
mining the literature

Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, Eivind Hovig.
A literature network of human genes for high throughput gene-expression analysis.
Nature Genetics, 28: 21-28
mining the literature

(at NTNU)

statistical methods
machine learning
natural language processing
challenges....

Information Bases/Derived-Data Databases
Experimental/Clinical Data

develop improved methods for modeling
modeling

data driven

first principles
**functional genomics**

**biological system**

- **generation of hypotheses**
  - biological roles of genes and proteins

- **screening**
  - genome-wide

- **modeling**
  - computational biology

- **experimental analysis**
  - functions of genes and proteins
problem

hypotheses

model

experiment/observation
multi-disciplinary effort

biological system

generation of hypoteses
biological roles of genes and proteins

metasciences

biology

medicine
genome-wide screening

technology

statistics

modeling
computational biology

computing

experimental analysis
functions of genes and proteins
Arne K. Sandvik
Helge L. Waldum
Fekadu Yadetie
Kristin Nørsett
Vidar Beisvåg
Berit Dosev Eitrem
Hallgeir Bergum
Frode Jünge
Torunn Bruland
Ola Ween
Liv Thommesen
Kristine Misund
Tonje Strømmen

Mette Langaas
Raymond Dingedine

Agnar Aamodt
Waclaw Kusnierczyk
Pauline Haddow
Gunnar Tufte
Kjell Bratbergsengen
Heri Ramampiaro
Tore Amble
Rune Sætre
Bjørn Alsberg
Arnar Flatberg
Lars Giskehaug
Torulf Mollestad
Henrik Tveit

Jan Komorowski
Torgeir Hvidsten
Herman Midelfart
Vladimir Yankovski

Norwegian Microarray Consortium: www.mikromatrise.no