Introduction to translational and clinical bioinformatics

Connecting complex molecular information to clinically relevant decisions using molecular profiles

Constantin F. Aliferis M.D., Ph.D., FACMI
Director, NYU Center for Health Informatics and Bioinformatics
Informatics Director, NYU Clinical and Translational Science Institute
Director, Molecular Signatures Laboratory,
Associate Professor, Department of Pathology,
Adjunct Associate Professor in Biostatistics and Biomedical Informatics, Vanderbilt University

Alexander Statnikov Ph.D.
Director, Computational Causal Discovery laboratory
Assistant Professor, NYU Center for Health Informatics and Bioinformatics, General Internal Medicine
Overview

• Session #1: Basic Concepts
• **Session #2: High-throughput assay technologies**
• Session #3: Computational data analytics
• Session #4: Case study / practical applications
• Session #5: Hands-on computer lab exercise
Molecular Signatures

**Definition** = computational or mathematical models that link high-dimensional molecular information to phenotype of interest

**Gene markers**

**New drug targets**
Molecular Signatures: Main Uses

1. Direct benefits: Models of disease phenotype/clinical outcome & estimation of the model performance
   - Diagnosis
   - Prognosis, long-term disease management
   - Personalized treatment (drug selection, titration) (“predictive” models)

2. Ancillary benefits 1: Biomarkers for diagnosis, or outcome prediction
   - Make the above tasks resource efficient, and easy to use in clinical practice
   - Helps next-generation molecular imaging
   - Leads for potential new drug candidates

3. Ancillary benefits 2: Discovery of structure & mechanisms (regulatory/interaction networks, pathways, sub-types)
   - Leads for potential new drug candidates
Less Conventional Uses of Molecular Signatures

• Increased Clinical Trial sample efficiency, and decreased costs or both, using placebo responder signatures;
• In silico signature-based candidate drug screening;
• Drug “resurrection”
• Establishing existence of biological signal in very small sample situations where univariate signals are too weak;
• Assess importance of markers and of mechanisms involving those
• Choosing the right animal model
• …?
Recent molecular signatures available for patient care
### Molecular signatures in the market (examples)

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Disease</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agendia</strong></td>
<td><strong>MammaPrint</strong></td>
<td>Breast cancer</td>
<td>Risk assessment for the recurrence of distant metastasis in a breast cancer patient.</td>
</tr>
<tr>
<td><strong>Agendia</strong></td>
<td><strong>TargetPrint</strong></td>
<td>Breast cancer</td>
<td>Quantitative determination of the expression level of estrogen receptor, progesteron receptor and HER2 genes. <em>This product is supplemental to MammaPrint.</em></td>
</tr>
<tr>
<td><strong>Agendia</strong></td>
<td><strong>CupPrint</strong></td>
<td>Cancer</td>
<td>Determination of the origin of the primary tumor.</td>
</tr>
<tr>
<td><strong>University Genomics</strong></td>
<td><strong>Breast Bioclassifier</strong></td>
<td>Breast cancer</td>
<td>Classification of ER-positive and ER-negative breast cancers into expression-based subtypes that more accurately predict patient outcome.</td>
</tr>
<tr>
<td><strong>Clarient</strong></td>
<td><strong>Insight Dx Breast Cancer Profile</strong></td>
<td>Breast cancer</td>
<td>Prediction of disease recurrence risk.</td>
</tr>
<tr>
<td><strong>Clarient</strong></td>
<td><strong>Prostate Gene Expression Profile</strong></td>
<td>Prostate cancer</td>
<td>Diagnosis of grade 3 or higher prostate cancer.</td>
</tr>
<tr>
<td><strong>Prediction Sciences</strong></td>
<td><strong>RapidResponse c-Fn Test</strong></td>
<td>Stroke</td>
<td>Identification of the patients that are safe to receive tPA and those at high risk for HT, to help guide the physician’s treatment decision.</td>
</tr>
<tr>
<td><strong>Genomic Health</strong></td>
<td><strong>OncotypeDx</strong></td>
<td>Breast cancer</td>
<td>Individualized prediction of chemotherapy benefit and 10-year distant recurrence to inform adjuvant treatment decisions in certain women with early-stage breast cancer.</td>
</tr>
<tr>
<td><strong>bioTheranostics</strong></td>
<td><strong>CancerTYPE ID</strong></td>
<td>Cancer</td>
<td>Classification of 39 types of cancer.</td>
</tr>
<tr>
<td><strong>bioTheranostics</strong></td>
<td><strong>Breast Cancer Index</strong></td>
<td>Breast cancer</td>
<td>Risk assessment and identification of patients likely to benefit from endocrine therapy, and whose tumors are likely to be sensitive or resistant to chemotherapy.</td>
</tr>
<tr>
<td><strong>Applied Genomics</strong></td>
<td><strong>MammaStrat</strong></td>
<td>Breast cancer</td>
<td>Risk assessment of cancer recurrence.</td>
</tr>
<tr>
<td><strong>Applied Genomics</strong></td>
<td><strong>PulmoType</strong></td>
<td>Lung cancer</td>
<td>Classification of non-small cell lung cancer into adenocarcinoma versus squamous cell carcinoma subtypes.</td>
</tr>
<tr>
<td><strong>Applied Genomics</strong></td>
<td><strong>PulmoStrat</strong></td>
<td>Lung cancer</td>
<td>Assessment of an individual's risk of lung cancer recurrence following surgery for helping with adjuvant therapy decisions.</td>
</tr>
<tr>
<td><strong>Correlogic</strong></td>
<td><strong>OvaCheck</strong></td>
<td>Ovarian cancer</td>
<td>Early detection of epithelial ovarian cancer.</td>
</tr>
<tr>
<td><strong>LabCorp</strong></td>
<td><strong>OvaSure</strong></td>
<td>Ovarian cancer</td>
<td>Assessment of the presence of early stage ovarian cancer in high-risk women.</td>
</tr>
<tr>
<td><strong>Veridex</strong></td>
<td><strong>GeneSearch BLN Assay</strong></td>
<td>Breast cancer</td>
<td>Determination of whether breast cancer has spread to the lymph nodes.</td>
</tr>
<tr>
<td><strong>Power3</strong></td>
<td><strong>BC-SeraPro</strong></td>
<td>Breast cancer</td>
<td>Differentiation between breast cancer patients and control subjects.</td>
</tr>
</tbody>
</table>
Molecular Signatures
Gene markers
New drug targets
An early kind of analysis: learning disease sub-types by clustering patient profiles
Clustering: seeking ‘natural’ groupings & hoping that they will be useful...
E.g., for treatment

Respond to treatment Tx1

Do not Respond to treatment Tx1
E.g., for diagnosis

\[ \text{Adenocarcinoma} \]

\[ \text{Squamous carcinoma} \]
Another use of clustering

• Cluster genes (instead of patients):
  – Genes that cluster together may belong to the same pathways
  – Genes that cluster apart may be unrelated
Unfortunately clustering is a non-specific method and falls into the ‘one-solution fits all’ trap when used for prediction.
Clustering is also non-specific when used to discover pathway membership, regulatory control, or other causation-oriented relationships.

It is entirely possible in this simple illustrative counter-example for G3 (a causally unrelated gene to the phenotype) to be more strongly associated and thus cluster with the phenotype (or its surrogate genes) more strongly than the true oncogenic genes G1, G2.
Brief overview of microarrays

Slides courtesy of: Stuart Brown, Ph.D.
Center for Health Informatics and Bioinformatics
Genomics

• Main array technologies (cDNA, Oligo, Tiled)

• Main uses:
  – Gene Expression
  – SNP assay
  – Gene copy number (array-CGH)
  – TF binding sites (Chip-on-chip)
  – Splice variation
What is a cDNA Microarray?

- **Hybridization – based** → put cDNA probes on an array surface and label the sample RNA
- Make probes for lots of genes - a massively parallel experiment
- Make it tiny so you don’t need so much RNA from your experimental cells.
- Make quantitative measurements
DNA Chip Microarrays

- Also hybridization based. Put a large number (~100K) of cDNA sequences or synthetic DNA oligomers onto a glass slide (or other substrate) in known locations on a grid.
- Label an RNA sample and hybridize
- Measure amounts of RNA bound to each square in the grid
- Make comparisons
  - Cancerous vs. normal tissue
  - Treated vs. untreated
  - Time course
cDNA Microarray Technologies

• Spot cloned cDNAs onto a glass microscope slide
  – usually PCR amplified segments of plasmids
• Label 2 RNA samples with 2 different colors of flourescent dye - control vs. experimental
• Mix two labeled RNAs and hybridize to the chip
• Make two scans - one for each color
• Combine the images to calculate ratios of amounts of each RNA that bind to each spot
Spot your own Chip
(plans available for free from Pat Brown’s website)

Robot spotter

Ordinary glass microscope slide
Micro-array Experiment

Prepare cDNA Probe
- Tumorigenic Cells
- Chromosome 6 Suppressed Cells
- Reverse Transcription
- Label with Fluorescent Dyes
- Combine Equal Amounts

Hybridize probe to microarray

SCAN
cDNA Spotted Microarrays
DNA Chip Microarrays

• Put a large number (~100K) of cDNA sequences or synthetic DNA oligomers onto a glass slide (or other substrate) in known locations on a grid.
• Label an RNA sample and hybridize
• Measure amounts of RNA bound to each square in the grid
• Make comparisons
  – Cancerous vs. normal tissue
  – Treated vs. untreated
  – Time course
• Many applications in both basic and clinical research
Affymetrix “Gene chip” system

- Uses 25 base oligos synthesized in place on a chip (~11 pairs of oligos for each gene)
- RNA labeled and scanned in a single “color” — one sample per chip
- Can have as many as 1,000,000 probes on a chip
- Arrays get smaller every year (more genes)
- Chips are ~expensive
- Proprietary system
Affymetrix Gene Chip

GeneChip expression analysis probe array

Image of hybridized probe array
Data Acquisition

• Scan the arrays
• Quantitate each spot
• Subtract background
• Normalize
• Export a table of fluorescent intensities for each gene in the array
Affymetrix Software

• Affymetrix System is totally automated
• Computes a single value for each gene from 40 probes - (using surprisingly kludgy math)
• Highly reproducible
  (re-scan of same chip or hyb. of duplicate chips with same labeled sample gives very similar results)
• Incorporates false results due to image artefacts
  – dust, bubbles
  – pixel spillover from bright spot to neighboring dark spots
• Visual Inspection: examples of defects in Affymetrix GeneChips
Upstream Basic Data Analysis

- Scan/quantitate/QC calls
- De-noise: Set cutoff filter for low values (background + noise)
- Normalize (i.e., remove measurement assay systematic biases)
- Fold change (relative increase or decrease in intensity for each gene) & log-transform
Brief overview of NGS (Next-Generation Sequencing)

Slides courtesy of: Zuojian Tang
Center for Health Informatics and Bioinformatics
Available next-generation sequencing platforms

- Illumina/Solexa
- Roche 454
- ABI SOLiD
- Polonator
- HeliScope
- ...
Next-Generation DNA Sequencing

• **Illumina** – 10 Million sequence reads per sample
  • Reads are 34-100 bp long
  • Paired End protocol is available

• **Roche/454** – 100 K sequences per sample
  • Reads are 250-450 bp long
  • Long Paired End protocol is available
Next-Gen Applications

• Genome whole sequencing
  • Sequence new genomes
  • re-sequence known genomes – find mutations/variation

• Microbiomics
  • Diversity and mutations of microbes and relationship to disease

• Targeted sequencing
  • PCR amplified regions

• ChIP-seq: identify regions of the genome bound by specific proteins
  • Transcription factors (promoter-based regulation)
  • Histone modification (epigenomics)

• RNA-seq: transcriptome sequencing
  • Gene expression (better sensitivity and accuracy than microarrays)
  • Mutations in coding sequences
  • Alternative splicing
  • miRNA
Strategies for cyclic array sequencing

(a) With the 454 platform, clonally amplified 28-m beads generated by emulsion PCR serve as sequencing features and are randomly deposited to a microfabricated array of picoliter-scale wells. With pyrosequencing, each cycle consists of the introduction of a single nucleotide species, followed by addition of substrate (luciferin, adenosine 5'-phosphosulphate) to drive light production at wells where polymerase-driven incorporation of that nucleotide took place. This is followed by an apyrase wash to remove unincorporated nucleotide.

(b) With the Solexa technology, a dense array of clonally amplified sequencing features is generated directly on a surface by bridge PCR. Each sequencing cycle includes the simultaneous addition of a mixture of four modified deoxynucleotide species, each bearing one of four fluorescent labels and a reversibly terminating moiety at the 3' hydroxyl position. A modified DNA polymerase drives synchronous extension of primed sequencing features. This is followed by imaging in four channels and then cleavage of both the fluorescent labels and the terminating moiety.
Illumina/Solexa Sequencer at NYU Medical Center
Sequencing Group

1. **SAMPLE PREPARATION**
   - ~6 hours (~3 hours hands-on)
   - Sample collection, genomic DNA sheared
   - DNA end-repair
   - Adapter ligation

2. **CLUSTER GENERATION**
   - ~5 hours
   - Flow cell and reagents placed into Cluster Station
   - Samples applied to flow cell
   - Complete walk-away automation

3. **SEQUENCING BY SYNTHESIS**
   - ~1 day (25 bp), ~5 days (2 x 50 bp)
   - Flow cell and reagents placed on Genome Analyzer
   - Complete walk-away automation

4. **PAIRED-END MODULE**
   - Add-on module for automated reagent delivery
   - Second read prepared and sequenced while flow cell remains on Genome Analyzer

Bioinformatics Group (Pre-stage Analysis)

- **RTA analysis (GA PC)**
  (2hrs longer than GAII Sequencing)
  - Image analysis
  - Base calling
  - Image transferring (IPAR)
  - Analysis results transferring (Cluster)

- **GA pipeline analysis (Cluster Server)**
  - Pipeline run (25hrs)
  - QC evaluation and update (1-5 hrs)
  - Compress and transfer data (5hrs)

- **Data archive (10hrs) (Cluster Sever)**
  - GA PC, IPAR, and cluster server space cleaning (2hrs)

- **Preliminary results (Cluster Server)**
  - Statistical summary (2hrs)
  - Parsing GA pipeline results (ChIP-Seq:5hrs)
  - Delivery preliminary results (2hrs)
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
Denaturation leaves single-stranded templates anchored to the substrate.

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase. After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

The data are aligned and compared to a reference, and sequencing differences are identified.
The Genome Analyzer Pipeline software is a highly customizable analysis engine capable of taking the raw image data generated by the Genome Analyzer and producing intensity scores, base calls and quality metrics, and quality scored alignments. It was developed in collaboration with many of the world’s leading sequencing centers and is scalable to meet the needs of even the most prodigious facilities.
Alignment and Polymorphism Detection

BFAST – Blat-like Fast Accurate Search Tool
Nils Homer, Stanley F. Nelson and Barry Merriman, University of California, Los Angeles
http://genome.ucla.edu/bfast

MAQ – Mapping and Assembly with Quality
Heng Li, Sanger Centre
http://maq.sourceforge.net/maq-man.shtml

Bowtie - An ultrafast memory-efficient short read aligner
Ben Langmead and Cole Trapnell, Center for Bioinformatics and Computational Biology, University of Maryland
http://bowtie-bio.sourceforge.net/
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Multi-threaded</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFAST</td>
<td>Explicit time and accuracy tradeoff with a prior accuracy estimation, supported by indexing the reference sequences. Optimally compresses indexes. Can handle billions of short reads. Can handle insertions, deletions, SNPs, and color errors (can map ABI SOLID color space reads). Performs a full Smith Waterman alignment.</td>
<td>Yes (POSIX)</td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>MAQ</td>
<td>Ungapped alignment that takes into account quality scores for each base.</td>
<td></td>
<td>GPL</td>
</tr>
<tr>
<td>Novoalign</td>
<td>Gapped alignment of single end and paired end Illumina GA I &amp; II reads. Uses base qualities. Reports can be converted to MAQ map files for analysis with MAQ utilities. Performance comparable to MAQ but with gapped alignment.</td>
<td>Multi-threading available with paid license</td>
<td>Single threaded version free for academic and non-commercial use.</td>
</tr>
<tr>
<td>ELAND</td>
<td>Implemented by Illumina. Includes ungapped alignment with a finite read length.</td>
<td></td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>SSAHA and</td>
<td>Fast for a small number of variants.</td>
<td></td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>SSAHA2</td>
<td></td>
<td></td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>SOAP</td>
<td>Robust with a small (1-3) number of gaps and mismatches. Speed improvement over BLAT, uses a 12 letter hash table.</td>
<td></td>
<td>GPL</td>
</tr>
<tr>
<td>BLAT</td>
<td>Made by Jim Kent. De facto standard for nucleotide sequence alignment, ungapped alignment, and can handle one mismatch in initial alignment step.</td>
<td>Yes (client/server).</td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>BLASTN</td>
<td>BLAST’s nucleotide alignment program, slow and not accurate for short reads, and uses a sequence database (EST, sanger sequence) rather than a reference genome.</td>
<td></td>
<td>Free for academic and non-commercial use.</td>
</tr>
<tr>
<td>ZOOM</td>
<td>Reads can vary between 25bp and 54bp in length. Fast but not accurate for &gt;2 mismatches. Ignores insertions and deletions.</td>
<td></td>
<td>Commercial</td>
</tr>
<tr>
<td>RMAP</td>
<td>Read lengths can range from 20bp to at most 64bp. Uses the &quot;exclusion principle&quot; to allow for mismatches and look-up reads in an index.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHRIMP</td>
<td>Indexes the reads instead of the reference genome. Uses masks to generate possible keys. Can map ABI SOLID color space reads.</td>
<td></td>
<td>BSD derivative</td>
</tr>
<tr>
<td>Bowtie</td>
<td>Uses a Burrows-Wheeler to create a permanent, reusable index of the genome; 1.3 GB memory footprint for human genome. Aligns more than 25 million Illumina reads in 1 CPU hour. Supports Maq-like and SOAP-like alignment policies.</td>
<td>Yes (POSIX)</td>
<td>Artistic License</td>
</tr>
<tr>
<td>QPalma</td>
<td>Uses known alignments to align targeted spliced reads. Useful for transcriptome resequencing and gene exploration.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOCS</td>
<td>For ABI SOLID technologies. Significant increase in time to map reads with mismatches (or color errors). Uses an iterative version of the Rabin-Karp string search algorithm.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MOSAIK</td>
<td>Fast and incorporates assembly. Aligns reads using a hashing scheme. Must split the reads many times to be robust against increasing number of mismatches.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SLIDER</td>
<td>SLIDER is an application for the Illumina Sequence Analyzer output that uses the &quot;probability&quot; files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Base calling

(a–d) A DNA cluster comprises identical DNA templates (colored boxes) that are attached to the flow cell. Nascent strands (black boxes) and DNA polymerase (black ovals) are depicted.
(a) In the ideal situation, after several cycles the signal (green arrows) is strong, coherent and corresponds to the interrogated position.
(b) Phasing noise introduces lagging (blue arrows) and leading (red arrow) nascent strands, which transmit a mixture of signals.
(c) Fading is attributed to loss of material that reduces the signal intensity (c).
(d) Changes in the fluorophore cross-talk cause misinterpretation of the received signal (teal arrows; d). For simplicity, the noise factors are presented separately from each other.
List of reasons caused bad results

- Library preparation:
  - Not enough DNA fragment
  - Not good adaptor attached
  - ...
- Sequencing problem:
  - Not good flow cell
  - Bubble problem
  - Oil spread problem
  - Not include Phix control problem
  - Inserted re-calibration problem
  - ...

Downstream data analysis for ChIP-Seq applications
Determining how proteins interact with DNA to regulate gene expression is essential for fully understanding many biological processes and disease states.

Specific DNA-protein interaction sites can be isolated by chromatin immunoprecipitation (ChIP). ChIP enriches for a library of target DNA sites that a given protein bound to in vivo.

Illumina combines whole-genome ChIP with massively parallel DNA sequencing to identify and quantify binding sites for DNA-associated proteins.

Illumina’s ChIP-Seq protocol cost-effectively and precisely maps global binding sites for a protein of interest across the entire genome.

The ChIP process enriches specific DNA-protein complexes using an antibody against a protein of interest.

Oligonucleotide adapters are then added to the small stretches of DNA that were bound to the protein of interest.

After size selection, the resulting ChIP DNA fragments are sequenced using the Cluster Station, Genome Analyzer, and Illumina Sequencing Reagents.

Low sample input requirements minimize tedious immunoprecipitations while comprehensive mapping across the whole genome deliver data at 1/10th to 1/30th the cost of conventional tilling array (ChIP-chip) experiments.

Most binding sites can be mapped using data generated in a single lane of one eight-lane flow cell.
Two common designs

• One sample experiment
  contains only a ChIP’d sample

• Two sample experiment
  contains a ChIP’d sample and a negative control sample (IgG/Input DNA)
ChIP-Seq application (Histone modification)

<table>
<thead>
<tr>
<th>lane</th>
<th>Sample Type</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>lane-1</td>
<td>Human</td>
<td>control h3k9</td>
</tr>
<tr>
<td>lane-2</td>
<td>Human</td>
<td>Ni h3k9</td>
</tr>
<tr>
<td>lane-3</td>
<td>Human</td>
<td>control h3k4</td>
</tr>
<tr>
<td>lane-4</td>
<td>Phix</td>
<td>Phix-Control</td>
</tr>
<tr>
<td>lane-5</td>
<td>Human</td>
<td>Ni h3k4</td>
</tr>
<tr>
<td>lane-6</td>
<td>Human</td>
<td>control input</td>
</tr>
<tr>
<td>lane-7</td>
<td>Human</td>
<td>control h3k4 (Non-Illumina's primers)</td>
</tr>
<tr>
<td>lane-8</td>
<td>Human</td>
<td>Ni h3k4 (non-Illumina's DNA library kit)</td>
</tr>
</tbody>
</table>

Purpose: the significant changes after nickel treatment
Lane-5 vs lane-3
Lane-8 vs lane-7
Question 8: ChIP-Seq vs ChIP-on-chip vs gene expression microarray
### ChIP-Seq application (Transcription Factor)

<table>
<thead>
<tr>
<th>Lane</th>
<th>Species</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mouse</td>
<td>IRF4 input</td>
</tr>
<tr>
<td>2</td>
<td>Mouse</td>
<td>IRF4 ChIP</td>
</tr>
<tr>
<td>3</td>
<td>Mouse</td>
<td>DP Input mono-nucleosomes</td>
</tr>
<tr>
<td>4</td>
<td>Phix</td>
<td>Phix-Control</td>
</tr>
<tr>
<td>5</td>
<td>Mouse</td>
<td>DP H3k4me3 ChIP condition 1</td>
</tr>
<tr>
<td>6</td>
<td>Mouse</td>
<td>DP H3k4me3 ChIP condition 2</td>
</tr>
<tr>
<td>7</td>
<td>Mouse</td>
<td>DP H3k36me3 ChIP condition 1</td>
</tr>
<tr>
<td>8</td>
<td>Mouse</td>
<td>DP H3k36me3 ChIP condition 2</td>
</tr>
</tbody>
</table>

**Purpose:** Find binding sites for IRF4
Lane-2 vs lane-1
## RNA-Seq application

<table>
<thead>
<tr>
<th>Lane</th>
<th>Sample Type</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>lane-1</td>
<td>Human</td>
<td>745580D</td>
</tr>
<tr>
<td>lane-2</td>
<td>Human</td>
<td>745580D</td>
</tr>
<tr>
<td>lane-3</td>
<td>Human</td>
<td>745580R</td>
</tr>
<tr>
<td>lane-4</td>
<td>Phix-Control</td>
<td>Phix</td>
</tr>
<tr>
<td>lane-5</td>
<td>Human</td>
<td>745580R</td>
</tr>
<tr>
<td>lane-6</td>
<td>Human</td>
<td>726584D</td>
</tr>
<tr>
<td>lane-7</td>
<td>Human</td>
<td>726584R</td>
</tr>
<tr>
<td>lane-8</td>
<td>???</td>
<td>Sample from Chris</td>
</tr>
</tbody>
</table>

6 Patient Samples with 3 runs totally  
First run is successful.  
Second run is questionable.  
Third run is not done yet.
Aims

2. New SNP, and D/R specific SNP category.
3. Fusion genes and new transcript.
4. D/R splicing form differential analysis
5. Correlate with microarray expression and CNV data.
Brief overview of proteomics & high throughput proteomic assays

Alexander Statnikov, Ph.D.
From Genomics to Proteomics

DNA

Transcription

mRNA

Translation

Protein
What is Proteome?

• A PROTEOME is the entire PROTein complement expressed by a genOMe, or by a cell or tissue type.

• There is only one definitive genome of an organism, the proteome is an entity which can change under different conditions, and can be dissimilar in different tissues of a single organism.

• The number of proteins in a proteome can exceed the number of genes present, as protein products expressed by alternative gene splicing or with different post-translational modifications.
Human Proteome

• Human genome: 30,000 genes

• Human proteome: ~ 1,000,000 protein variants??
  – alternative splicing, co-/post-translational modifications, post-translational processing etc. add complexity to the human proteome

• Genome – Static
• Proteome – Very Dynamic
Definition of Proteomics

- Proteomics is the study of total protein complements, proteomes, e.g. from a given tissue or cell type.
Why Proteomics?

mRNA level (transcriptome) does not always reflect protein expression level

• Comparative expression analyses of mRNA and proteins have shown that expression levels of mRNAs are not necessarily correlated with those of the encoded proteins.
  – Different stability of mRNAs
  – Different protein translation rate
  – Different half-lives of proteins
  – Post-translational processing/modification
Goal of Proteomics

• Quantitatively characterize all proteins expressed by a tissue or organism
  – Complete protein expressions
  – Complete covalent structures
  – Complete protein networks
  – Complete 3D structures
  – Complete functional assignments

• Understanding how proteins function in the living cells
Tools to Study Proteomics
Proteomics: Analytical Challenges

- **Proteome**
  - Very dynamic
  - Very complex
  - A huge range of protein abundances
  - Variable solubility
  - Can not be amplified
Proteomics Needs

– High resolution protein/peptide separation techniques
  • Electrophoresis
  • High Performance Liquid Chromatography

– Highly sensitive detection techniques
  • Mass spectrometry

– Reliable bioinformatics tools
Basics of Proteomics

A protein sample is digested (typically with \textbf{trypsin}) to generate \textbf{peptides}.

The peptides are then separated by liquid chromatography.
Basics of Proteomics

The mass spectrometer separates the eluting peptides by mass-to-charge ratio (m/z), and records a **mass spectrum.**
General Flow Scheme for Proteomic Analysis

**Top-down method**

Protein mixture → 2D-PAGE → Proteins

Protein mixture → Digestion → Peptide mixture → HPLC → Peptides

**Bottom-up/Shotgun methods**

Peptide mixture → Digestion → Electrospray/MALDI/SELDI → MS analysis → MS data
How does Mass Spectrometer operate?

Ionization  
Ion Separation  
Ion Detection
Mass spectrometers generate charged species (e.g. molecular ion) and then sort them based on mass-to-charge ($m/z$) ratio.
Above: Diagram of a mass spectrometer (courtesy of ChemGuide.com).

Molecules are accelerated by a series of charged plates, their time of flight determined by their mass-to-charge ratio.
Mass Spectrometry Techniques Used in Analysis of Peptides and Proteins

Mass Spectrometers are usually classified on the basis of how samples are ionized and how the mass separation is accomplished.

• **Ionization techniques**
  – Electrospray ionization (ESI), Nano-ESI
  – Matrix-assisted laser desorption ionization (MALDI)
  – Surface enhanced laser desorption ionization (SELDI)

• **Mass analyzers**
  – Quadrupole
  – Time-of-flight
  – Ion trap
  – Quadrupole/quadrupole
  – Quadrupole/Time-of-flight
  – Quadrupole/Ion trap
  – Time-of-flight/Time-of-flight
  – Ion-trap/Fourier-transform ion cyclotron resonance
ESI

MALDI
How MS Data Looks Like?

• Spectra produced
  – **Mass/charge** ratio (m/z) plotted against **relative intensity**
  – $10^4$ - $10^6$ data points per spectrum
  – Sample SELDI-TOF spectrum:
Basics of Tandem MS

Secondary Fragmentation

Ionized parent peptide
Peptide Sequencing by MS/MS

MGLAAAVFTK

Survey scan (mass spectrum) (MS1)

q1

q2

TOF

Tandem mass spectrum (MS2)

On-line separation

GLAAAVFTK^+

AAVFTK^+

LAAAVFTK^+

FTK^+

K^+
Protein Identification MS/MS data

![MS/MS spectrum diagram]

- Relative Intensity (%)
- m/z values: 200, 400, 600, 800, 1000, 1200, 1400...
- Intensity (%) values: 100, 50, 20, 10, 5, 2, 1
- Peptide sequence: LGEnYGFQNALIVR
- B and Y ions: b\textsubscript{2}, b\textsubscript{3}, b\textsubscript{4}, y\textsubscript{1}, y\textsubscript{2}, y\textsubscript{3}, y\textsubscript{4}, y\textsubscript{5}, y\textsubscript{6}, y\textsubscript{7}, y\textsubscript{8}, y\textsubscript{9}, y\textsubscript{10}, y\textsubscript{11}, y\textsubscript{12}
- Mass values: 171.13, 175.12, 300.16, 500.22, 500.36, 571.40, 685.47, 813.53, 960.62, 1017.62, 1225.67, 1354.74, 1411.77
Protein Identification MS/MS data

Bioinformatics tools (database search software) is used

Proteins in the data base
Trypsin → Tryptic peptides

Protein of interest
Trypsin → Tryptic peptides

LCESI-MS/MS

Protein ID

m/z

Proteomics relies on genome sequence databases!!!
Importance of Mass Spectrometry Data

- Detection of a disease

  - Promising results on various types of cancer

- Discovery of disease biomarkers
  - Identify species (protein/peptides) responsible for the differences
Challenges in MS data analysis

Many sources of variation
- Sample collection, sample storage & sample processing
- Instrument conditions

+ natural variation in protein expression levels among individuals

Example: profiles of two patients with pancreatic cancer
Challenges in data analysis

Three types of systematic instrument errors

- Baseline shift
- Intensity measurement error
- Mass inaccuracy

Example: two profiles for the same reference serum
Profile preprocessing

- Aims is to remove the noise and systematic biases in the signal while preserving useful information
- Typical preprocessing steps:
  - **Smoothing:**
    - Aims to eliminate the noise in the signal
  - **Calibration/rescaling:**
    - Attempts to eliminate differences in intensity among profiles
  - **Profile transformations (variance reduction):**
    - Reduces the multiplicative noise
  - **Baseline correction:**
    - Reduces the systematic baseline error

Peak detection
Peak alignment
Preprocessing: Smoothing

- **Aims**: to eliminate a high-frequency noise in the signal
- **Threat**: a loss of information in the high frequency signal
Preprocessing: Baseline correction

- **Aims is to** eliminate a systematic intensity bias by which the profile readings differ from 0.
Comparison to Microarrays

**Microarray Analysis**

- Do these represent the same peak?

**Mass Spectrometry Analysis**

- Correspond to the same probe and designed to have, ideally, one-to-one correspondence to the same gene.
- How many peptides does this grouping represent?
- Is this a peak?
Comparison to Microarrays

1. Microarray analysis of nucleotides
   - All spots may be known *a priori*
   - The array is the same (spots are “aligned”) from sample to sample
   - Intensity represents extent of hybridization with known oligonucleotides
   - Possible to limit analysis to known physiologic/pathologic pathways

2. Mass spectrometry analysis of peptides
   - Peptides represented by peaks are not known *a priori*
     - *A peak may represent: noise, single peptide (known or unknown), peptide amalgamation*
   - M/Z values are not aligned from sample to sample
   - Peak alignment is not straight-forward
   - Not possible to limit analysis to known physiologic/pathologic pathways
   - Spectra may represent tens to hundreds of thousands of data points
   - Lack of software performing complete analysis
Downstream analysis:
Two improved classes of methods (over clustering)

• Supervised learning $\rightarrow$ predictive signatures and markers
• Regulatory network reverse engineering $\rightarrow$ pathways
Supervised learning: use the known phenotypes (a.k.a “labels”) in training data to build signatures or find markers highly specific for that phenotype.
Regulatory network reverse engineering
Supervised learning: a geometrical interpretation

Cancer patients

New case, classified as cancer

New case, classified as normal

SVM classifier

Normals
In 2-D looks good but what happens in:

- 10,000-50,000 (regular gene expression microarrays, aCGH, and early SNP arrays)
- >500,000 (tiled microarrays, new SNP arrays)
- 10,000-300,000 (regular MS proteomics)
- >10,000,000 (LC-MS proteomics)

This is the ‘curse of dimensionality problem’
High-dimensionality (especially with small samples) causes:

• Some methods do not run at all (classical regression)
• Some methods give bad results (KNN, Decision trees)
• Very slow analysis
• Very expensive/cumbersome clinical application
• Tends to “overfit”
Two (very real and very unpleasant) problems: Over-fitting & Under-fitting

- **Over-fitting (a model to your data)** = building a model that is good in original data but fails to generalize well to fresh data
- **Under-fitting (a model to your data)** = building a model that is poor in both original data and fresh data
Intuitive explanation of overfitting & underfitting

• Play the game: find rule to predict who are the instructors in any given class (use today’s class to find a general rule)
Over/under-fitting are directly related to the complexity of the decision surface and how well the training data is fit.
Over/under-fitting are directly related to the complexity of the decision surface and how well the training data is fit
Very Important Concept:

• Successful data analysis methods balance training data fit with complexity.
  – Too complex signature (to fit training data well) \(\rightarrow\) overfitting (i.e., signature does not generalize)
  – Too simplistic signature (to avoid overfitting) \(\rightarrow\) underfitting (will generalize but the fit to both the training and future data will be low and predictive performance small).
Part of the Solution: feature selection
How well supervised learning works in practice?
Datasets

- Bhattacharjee2 - Lung cancer vs normals [GE/DX]
- Bhattacharjee2_I - Lung cancer vs normals on common genes between Bhattacharjee2 and Beer [GE/DX]
- Bhattacharjee3 - Adenocarcinoma vs Squamous [GE/DX]
- Bhattacharjee3_I - Adenocarcinoma vs Squamous on common genes between Bhattacharjee3 and Su [GE/DX]
- Savage - Mediastinal large B-cell lymphoma vs diffuse large B-cell lymphoma [GE/DX]
- Rosenwald4 - 3-year lymphoma survival [GE/CO]
- Rosenwald5 - 5-year lymphoma survival [GE/CO]
- Rosenwald6 - 7-year lymphoma survival [GE/CO]
- Adam - Prostate cancer vs benign prostate hyperplasia and normals [MS/DX]
- Yeoh - Classification between 6 types of leukemia [GE/DX-MC]
- Conrads - Ovarian cancer vs normals [MS/DX]
- Beer_I - Lung cancer vs normals (common genes with Bhattacharjee2) [GE/DX]
- Su_I - Adenocarcinoma vs squamous (common genes with Bhattacharjee3) [GE/DX]
- Banez - Prostate cancer vs normals [MS/DX]
Methods: Gene and Peak Selection Algorithms

- **ALL** - No feature selection
- **LARS** - LARS
- **HITON_PC**
- **HITON_PC_W** - HITON_PC+ wrapping phase
- **HITON_MB**
- **HITON_MB_W** - HITON_MB + wrapping phase
- **GA_KNN** - GA/KNN
- **RFE** - RFE with validation of feature subset with optimized polynomial kernel
- **RFE_Guyon** - RFE with validation of feature subset with linear kernel (as in Guyon)
- **RFE_POLY** - RFE (with polynomial kernel) with validation of feature subset with polynomial optimized kernel
- **RFE_POLY_Guyon** - RFE (with polynomial kernel) with validation of feature subset with linear kernel (as in Guyon)
- **SIMCA** - SIMCA (Soft Independent Modeling of Class Analogy): PCA based method
- **SIMCA_SVM** - SIMCA (Soft Independent Modeling of Class Analogy): PCA based method with validation of feature subset by SVM
- **WFCCM_CCR** - Weighted Flexible Compound Covariate Method (WFCCM) applied as in Clinical Cancer Research paper by Yamagata (analysis of microarray data)
- **WFCCM_Lancet** - Weighted Flexible Compound Covariate Method (WFCCM) applied as in Lancet paper by Yanagisawa (analysis of mass-spectrometry data)
- **UAF_KW** - Univariate with Kruskal-Wallis statistic
- **UAF_BW** - Univariate with ratio of genes between groups to within group sum of squares
- **UAF_S2N** - Univariate with signal-to-noise statistic
Classification Performance
(average over all tasks/datasets)

<table>
<thead>
<tr>
<th>Area under ROC curve</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>86.43%</td>
</tr>
<tr>
<td>LARS*</td>
<td>83.06%</td>
</tr>
<tr>
<td>HITONgp_PC</td>
<td>83.89%</td>
</tr>
<tr>
<td>HITONgp_MB</td>
<td>83.40%</td>
</tr>
<tr>
<td>HITONgp_PC_W</td>
<td>82.84%</td>
</tr>
<tr>
<td>HITONgp_MB_W</td>
<td>82.12%</td>
</tr>
<tr>
<td>GA_KNN*</td>
<td>83.58%</td>
</tr>
<tr>
<td>RFE</td>
<td>83.46%</td>
</tr>
<tr>
<td>RFE_Guyon</td>
<td>83.39%</td>
</tr>
<tr>
<td>RFE_POLY</td>
<td>83.26%</td>
</tr>
<tr>
<td>RFE_POLY_Guyon</td>
<td>83.67%</td>
</tr>
<tr>
<td>SIMCA</td>
<td>80.51%</td>
</tr>
<tr>
<td>SIMCA_SVM</td>
<td>85.11%</td>
</tr>
<tr>
<td>WFCCM_CCR*</td>
<td>83.91%</td>
</tr>
<tr>
<td>UAF_KW</td>
<td>85.02%</td>
</tr>
<tr>
<td>UAF_BW</td>
<td>86.32%</td>
</tr>
<tr>
<td>UAF_S2N</td>
<td>84.90%</td>
</tr>
</tbody>
</table>
How well gene selection works in practice?
Number of Selected Features (average over all tasks/datasets)
Number of Selected Features (zoom on most powerful methods)
Number of Selected Features (average over all tasks/datasets)

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>9963.30</td>
</tr>
<tr>
<td>LARS</td>
<td>194.58</td>
</tr>
<tr>
<td>HITONgp_PC</td>
<td>14.88</td>
</tr>
<tr>
<td>HITONgp_MB</td>
<td>27.53</td>
</tr>
<tr>
<td>HITONgp_PC_W</td>
<td>6.97</td>
</tr>
<tr>
<td>HITONgp_MB_W</td>
<td>8.99</td>
</tr>
<tr>
<td>GA_KNN</td>
<td>50.00</td>
</tr>
<tr>
<td>RFE</td>
<td>58.37</td>
</tr>
<tr>
<td>RFE_Guyon</td>
<td>69.34</td>
</tr>
<tr>
<td>RFE_POLY</td>
<td>97.65</td>
</tr>
<tr>
<td>RFE_POLY_Guyon</td>
<td>106.92</td>
</tr>
<tr>
<td>SIMCA</td>
<td>3060.31</td>
</tr>
<tr>
<td>SIMCA_SVM</td>
<td>1312.14</td>
</tr>
<tr>
<td>WFCCM_CCR</td>
<td>646.34</td>
</tr>
<tr>
<td>UAF_KW</td>
<td>2461.00</td>
</tr>
<tr>
<td>UAF_BW</td>
<td>4379.93</td>
</tr>
<tr>
<td>UAF_S2N</td>
<td>3224.59</td>
</tr>
</tbody>
</table>
Conclusions so far

• Special classifiers (with inherent complexity control) combined with feature selection & careful parameterization protocols overcome over-fitting & estimate future performance accurately.

• Caveats: analysis is typically complex and error prone. Need: (a) an experienced analyst on the team, or (b) a validated software system designed for non-experts.
Software

- Causal Explorer
- Gems
- Fast-aims
Causal Explorer

• Matlab library of computational causal discovery and variable selection algorithms
• Introductory-level library to our causal algorithms (~3% of our algorithms)
• Discover the direct causal or probabilistic relations around a response variable of interest (e.g., disease is directly caused by a set of variables/observed quantities).
• Discover the set of all direct causal or probabilistic relations among the variables.
• Discover the Markov blanket of a response variable of interest, i.e., the minimal subset of variables that contains all necessary information to optimally predict the response variable.
• Code emphasizes efficiency, scalability, and quality of discovery
• Requires relatively deep understanding of underlying theory and how the algorithms operate
Statistics of Registered Users

- **739** registered users in >50 countries.
- **402** (54%) users are affiliated with educational, governmental, and non-profit organizations.
- **337** (46%) users are either from private or commercial sectors.
- Major commercial organizations that have registered users of *Causal Explorer* include:
  - IBM
  - Intel
  - SAS Institute
  - Texas Instruments
  - Siemens
  - GlaxoSmithKline
  - Merck
  - Microsoft
## Statistics of Registered Users

Major U.S. institutions that have registered users of *Causal Explorer*:

<table>
<thead>
<tr>
<th>Boston University</th>
<th>Medical College of Wisconsin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brandies University</td>
<td>Michigan State University</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>Naval Postgraduate School</td>
</tr>
<tr>
<td>Case Western Reserve University</td>
<td>New York University</td>
</tr>
<tr>
<td>Central Washington University</td>
<td>Northeastern University</td>
</tr>
<tr>
<td>College of William and Mary</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Cornell University</td>
<td>Oregon State University</td>
</tr>
<tr>
<td>Duke University</td>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td>Harvard University</td>
<td>Princeton University</td>
</tr>
<tr>
<td>Illinois Institute of Technology</td>
<td>Rutgers University</td>
</tr>
<tr>
<td>Indiana University-Purdue University Indianapolis</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>State University of New York</td>
</tr>
<tr>
<td>Louisiana State University</td>
<td>Tufts University</td>
</tr>
<tr>
<td>M. D. Anderson Cancer Center</td>
<td>University of Arkansas</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
<td>University of California Berkeley</td>
</tr>
<tr>
<td>Medical College of Wisconsin</td>
<td>University of California Los Angeles</td>
</tr>
<tr>
<td>Michigan State University</td>
<td>University of California San Diego</td>
</tr>
<tr>
<td>Naval Postgraduate School</td>
<td>University of California Santa Cruz</td>
</tr>
<tr>
<td>New York University</td>
<td>University of Cincinnati</td>
</tr>
<tr>
<td>Northeastern University</td>
<td>University of Colorado Denver</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>University of Delaware</td>
</tr>
<tr>
<td>Oregon State University</td>
<td>University of Houston-Clear Lake</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>University of Idaho</td>
</tr>
<tr>
<td>Princeton University</td>
<td>University of Illinois at Chicago</td>
</tr>
<tr>
<td>Rutgers University</td>
<td>University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>Stanford University</td>
<td>University of Kansas</td>
</tr>
<tr>
<td>State University of New York</td>
<td>University of Maryland Baltimore County</td>
</tr>
<tr>
<td>Tufts University</td>
<td>University of Massachusetts Amherst</td>
</tr>
<tr>
<td>University of Arkansas</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>University of California Berkley</td>
<td>University of New Mexico</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>University of California San Diego</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>University of California Santa Cruz</td>
<td>University of Rochester</td>
</tr>
<tr>
<td>University of Cincinnati</td>
<td>University of Tennessee Chattanooga</td>
</tr>
<tr>
<td>University of Colorado Denver</td>
<td>University of Texas at Austin</td>
</tr>
<tr>
<td>University of Delaware</td>
<td>University of Utah</td>
</tr>
<tr>
<td>University of Houston-Clear Lake</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>University of Idaho</td>
<td>University of Washington</td>
</tr>
<tr>
<td>University of Illinois at Chicago</td>
<td>University of Wisconsin-Madison</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>University of Wisconsin-Milwaukee</td>
</tr>
<tr>
<td>University of Kansas</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>University of Maryland Baltimore County</td>
<td>Virginia Tech</td>
</tr>
<tr>
<td>University of Massachusetts Amherst</td>
<td>Yale University</td>
</tr>
</tbody>
</table>
Other systems for supervised analysis of microarray data

<table>
<thead>
<tr>
<th>Name</th>
<th>Version</th>
<th>Developer</th>
<th>Automatic model selection for classifier and gene selection methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayMiner ClassMarker</td>
<td>5.2</td>
<td>Optimal Design, Belgium</td>
<td>No</td>
</tr>
<tr>
<td>Avadis Prophetic</td>
<td>3.3</td>
<td>Strand Genomics, USA</td>
<td>No</td>
</tr>
<tr>
<td>BRB ArrayTools</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>caGEDA</td>
<td>1.0</td>
<td>Stanford University, USA</td>
<td>No</td>
</tr>
<tr>
<td>Cleaver</td>
<td>1.02</td>
<td>Applied Maths, Belgium</td>
<td>No</td>
</tr>
<tr>
<td>GeneCluster2</td>
<td>2.1.7</td>
<td>Broad Institute, Massachusetts Institute of Technology, USA</td>
<td>No</td>
</tr>
<tr>
<td>GeneLinker Platinum</td>
<td>4.5</td>
<td>Predictive Patterns Software, Canada</td>
<td>No</td>
</tr>
<tr>
<td>GeneMaths XT</td>
<td>1.02</td>
<td>Applied Maths, Belgium</td>
<td>No</td>
</tr>
<tr>
<td>GenePattern</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genesis</td>
<td>1.5.0</td>
<td>Graz University of Technology, Austria</td>
<td>No</td>
</tr>
<tr>
<td>GeneSpring</td>
<td>1.5</td>
<td>National Center for Cancer Research (CNIO), Spain</td>
<td>No</td>
</tr>
<tr>
<td>GEPAS</td>
<td>1.1</td>
<td>National Center for Cancer Research (CNIO), Spain</td>
<td>Limited (for number of genes)</td>
</tr>
<tr>
<td>MultiExperiment Viewer</td>
<td>3.0.3</td>
<td>The Institute for Genomic Research, USA</td>
<td>No</td>
</tr>
<tr>
<td>PAM</td>
<td>1.21a</td>
<td>Stanford University, USA</td>
<td>Limited (for a single parameter of the classifier)</td>
</tr>
<tr>
<td>Partek Predict</td>
<td>6.0</td>
<td>Partek, USA</td>
<td>Limited (does not allow optimization of the choice of gene selection algorithms)</td>
</tr>
<tr>
<td>Weka Explorer</td>
<td>3.4.3</td>
<td>University of Waikato, New Zealand</td>
<td>No</td>
</tr>
</tbody>
</table>

There exist many good software packages for supervised analysis of microarray data, but…

- Neither system provides a protocol for data analysis that precludes overfitting.
- A typical software either offers an overabundance of algorithms or algorithms with unknown performance.
- The software packages address needs only of experienced analysts.
Purpose of GEMS

Gene expression data and outcome variable

Normal  Cancer  Cancer  Normal  Normal  Cancer  ...
Cancer  Cancer  Normal

Optional: Gene names & IDs

ring finger protein 1
tubulin, beta, 5
glucose-6-phosphate dehydrogenase
glutathione S-transferase M5
carnitine acetyltransferase
Rho GTPase activating protein 4
SMA3
mannose phosphate isomerase
mitogen-activated protein kinase 3
leukotriene A4 hydrolase
chromosome 21 open reading frame 1
dihydropyrimidinase-like 2
beta-2-microglobulin
discs, large (Drosophila) homolog 4

Classification model

Cross-validation performance estimate

Reduced set of genes
Rho GTPase activating protein 4
SMA3
mannose phosphate isomerase
mitogen-activated protein kinase 3

Links to literature

(model generation & performance estimation mode)
Purpose of GEMS

Gene expression data and unknown outcome variable

Classification model


Model predictions

Normal
Cancer
Normal
Cancer
...
Cancer
Normal

Performance estimate

(model application mode)
Methods Implemented in GEMS

Cross-Validation Designs
- N-Fold CV
- LOOCV

Classifiers
- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS

Gene Selection Methods
- S2N One-Versus-Rest
- S2N One-Versus-One
- Non-param. ANOVA
- BW ratio
- HITON_MB
- HITON_PC

Normalization Techniques
- [a, b]
- \( \frac{x - \text{MEAN}(x)}{\text{STD}(x)} \)
- \( \frac{x}{\text{STD}(x)} \)
- \( \frac{x}{\text{MEAN}(x)} \)
- \( \frac{x}{\text{MEDIAN}(x)} \)
- \( \frac{x}{\text{NORM}(x)} \)
- \( \frac{x - \text{MEAN}(x)}{\text{MEDIAN}(x)} \)
- \( \frac{x}{\text{MEDIAN}(x)} \)
- \( \frac{x}{\text{NORM}(x)} \)
- ABS(x)
- x + ABS(x)

Performance Metrics
- Accuracy
- RCI
- AUC ROC
Software Architecture of GEMS

GEMS 2.0

Wizard-Like User Interface

Computational Engine

- Estimate classification performance
- Generate a classification model and estimate its performance
- Generate a classification model
- Apply existing model to a new set of patients

Cross-Validation Loop for Performance Est.
- N-Fold CV
- LOOCV

Cross-Validation Loop for Model Selection
- N-Fold CV
- LOOCV

Performance Computation
- Accuracy
- RCI
- AUC ROC

Normalization
- Gene Selection
  - $S2N$ One-Versus-Rest
  - $S2N$ One-Versus-One
  - Non-param. ANOVA
  - BW ratio
  - HITON_PC
  - HITON_MB

Classification by MC-SVM
- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS

Report Generator
GEMS 2.0: Wizard-Like Interface

- Task selection
- Dataset specification
- Cross-validation design
- Normalization
- Logging
- Performance metric
- Gene selection
- Classification
- Report generation
- Analysis execution
GEMS 2.0: Wizard-Like Interface

Input microarray gene expression dataset

File with gene names

File with gene accession numbers

Output model

---

GEMS: Experimental Report

- Task: Generate best model
- Experiment execution time: 46 seconds
- Number of samples: 203
- Number of variables: 12601
- Number of categories: 5
- Validation accuracy: 96.55176
- Dataset filename: DSashaMatlabNMCSVMToolbox/Development/distributive/data/Lung_Cancer-data.mat
- Gene names filename: DSashaMatlabNMCSVMToolbox/Development/distributive/data/Lung_Cancer_gene_names.txt
- Gene accession numbers filename: DSashaMatlabNMCSVMToolbox/Development/distributive/data/Lung_Cancer_gene_accessions.txt
- Model filename: model_mod

Description of the best model for the current data-split:
- SVM method: OVR
- SVM cost: 100
- SVM kernel poly
- SVM kernel parameter (degree): 1

Feature selection method: Signal-to-noise ratio in a one-versus-rest fashion

<table>
<thead>
<tr>
<th>Ranking (1-best)</th>
<th>Column index of features (in dataset file)</th>
<th>Gene names</th>
<th>Accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8458</td>
<td>Cluster 1185861: Human protein tyrosine phosphatase receptor P1 (PTPRB) mRNA, complete cds</td>
<td>UB1561</td>
</tr>
<tr>
<td>2</td>
<td>5850</td>
<td>E74 interacting protein B</td>
<td>AF055026</td>
</tr>
<tr>
<td>3</td>
<td>3074</td>
<td>picopicotyrosine/retinoblastoma protein</td>
<td>AB011123</td>
</tr>
<tr>
<td>4</td>
<td>8473</td>
<td>Clustering 149437: Human amyloid precursor-like protein 1 mRNA, complete cds</td>
<td>UB04677</td>
</tr>
<tr>
<td>5</td>
<td>4854</td>
<td>Clustering 1190862:zbl1062:1 Homo sapiens cDNA, complete cds</td>
<td>L26703</td>
</tr>
<tr>
<td>6</td>
<td>3876</td>
<td>cadherin, EGF LAG seven-pass G-type receptor 3, flanking (Drosophila) homolog</td>
<td>AB011536</td>
</tr>
<tr>
<td>7</td>
<td>3152</td>
<td>$100 calcium-binding protein A11 (caligataxin)</td>
<td>D38583</td>
</tr>
</tbody>
</table>
Statistics of registered users

- **800** users in **>50** countries
- **350** academic & non-profit users
- **450** private & commercial users
- **205** scientific citations of major paper that introduced GEMS
- Major commercial organizations that have registered users of *Causal Explorer* include:
  - Eli Lilly
  - IBM
  - Genedata
  - GenomicTree
  - Pronota
  - Novartis
  - GE
  - Nuvera Biosciences
  - Cogenetics
FAST-AIMS

- FAST-AIMS is a system to support automatic development of high-quality classification models and biomarker discovery in mass spectrometry proteomics data
- Incorporates automated data analysis protocols of GEMS
- Deals with additional challenges of MS data analysis
System Workflow

1. Get Mass Spectra
2. Data Pre-Processing
   - Range Restriction
   - Subtract Baseline
   - Noise Reduction
   - Normalization
   - Peak Extraction
   - Peak Alignment
3. Build Model
4. Apply Model
   - Select Train and Test Sets
     - Feature Selection
     - Peak Alignment
     - Classifier(s) Optimization
     - Cross-Validation
   - 1-fold n-fold LOOCV
   - Classify samples
   - Rank classification
   - Apply Performance Metric
   - ROC Sensitivity/Specificity Accuracy
Evaluation in multiple user study

**Figure 3: Performance of Generated Models**

- Expert
- Non-expert
- Biostatistician

**Figure 4: Performance comparison between current study and previous study**

- FAST-AIMS
- Biostatistician
- Weak cation exchange array
- Copper metal affinity capture array

+ Model developed independently of FAST-AIMS
++ Range bars represent ROC range (minimum to maximum)
Main points, session #2

- We reviewed basic principles of major high-throughput molecular assays:
  - Mass spectrometry proteomics
  - Microarrays
  - Next generation sequencing
- We continued the introduction of basic computational concepts such as over and under fitting, and dimensionality reduction.
- Next session: we will go deeper in computational techniques and related pitfalls.
For session #3

• Review slides in today’s presentation and bring written questions (if any) to discuss in subsequent sessions