Biomarkers for personalized medicine

Lars Arvastson
H. Lundbeck A/S

Danish pharma – still doing well!

INTRODUCING LUNDBECK

LUNDBECK IN BRIEF

We are an international pharmaceutical company specializing in central nervous system disorders

- → Founded by Hans Lundbeck in 1915
- → An integrated company with core competencies in research, development, production, marketing and sales
- → International presence with pharmaceuticals in more than 100 markets
- → Marketed pharmaceuticals include treatments for Alzheimer's disease, depression and anxiety, epilepsy, Huntington's disease, insomnia, Parkinson's disease, and schizophrenia/bipolar disorder
- → Headquarters in Copenhagen, Denmark
- → Approximately 6,000 employees in 57 countries
- → 2011 revenue: DKK 16 billion (approx. EUR 2.1 billion/USD 3 billion)

Biomarkers and personalized medicine

- → The future will be more focused on personalized treatment
- → Biomarkes have a central role
  - Diagnostic biomarkers
  - Prognostic biomarkers
  - Predictive biomarkers
  - Biological understanding

Molecular biomarkers – what to measure?

DNA – Gene expression – Proteins

State – Trait

Genotype – Phenotype

Cost – Quality

CNS – Periphery

Explorative – Hypothesis

Gene expression analysis

- → Genome wide scan
  - Micro array technology
  - ~100 000 genes
  - Low quality data
  - No prior assumptions
- → Selected candidate genes
  - qPCR technology
  - ~100 genes
  - High quality data
  - Selected based on prior knowledge

Scientific questions

- → Genes associated with a disease
  - Biological understanding
- → Prediction of treatment response
  - Companion diagnostic
- → Classification of disease state
  - Diagnosis

Associating genes with a diagnosis

$$x_{ij} = \mu_j + \beta_j^{\text{Age}} \text{Age}_i + \beta_j^{\text{Gender}} 1_{\{\text{Subj}_i = \text{Female}\}} + \beta_j^{\text{MDD}} 1_{\{\text{Subj}_i \text{ has MDD}\}} + \varepsilon_{ij}$$

$$x_{ij} = \text{Gene expression level for gene } j, \text{subject } i$$

- → Simple t-test
- → Multiple testing
  - Bonferroni
  - False Discovery Rate
- → Other confounding factors
  - BMI
  - Inclusion criteria
  - Smoking
  - Alcohol

Associating genes with treatment response

$$y_i = \mu + \sum_g \beta_g x_{gi} + \sum_g \gamma_g x_{gi} 1_{\{\text{Active treatement of subject}\, i\}} + \varepsilon_i$$

where

y_i is the outcome depression score (adjusted for baseline score and treatment),

x_{gi} is the gene expression for gene g, patient i,

$\varepsilon_i \sim IID\ N(0, \sigma^2)$

Predictive index for patient $i : P_i(\mathbf{x}) = \sum_g \hat{\gamma}_g \cdot x_{gi}$

Classification of disease status based on gene expresssion

Logistic model

$$\text{P(Subject } i \text{ has MDD} \mid x_{i1}, \ldots, \text{x}_{\text{in}}) = \frac{\exp\left(\beta_0 + \sum_{g=1}^{n} \beta_g x_{ig}\right)}{1 + \exp\left(\beta_0 + \sum_{g=1}^{n} \beta_g x_{ig}\right)}$$

$$x_{ij} = \text{Gene expression level for gene } j, \text{subject } i$$

Mathematical toolbox

- → Low dimensional data an a lot of data
  - Standard statistical toolbox
- → High dimensional data and few data
  - Regularization
    - $\min_\beta \left(\|y - \hat{y}(\beta \mid x)\|_2^2 + \lambda \|\beta\|_p \right)$
    - $\min_\beta \left(-L(\beta \mid X) + \lambda \|\beta\|_p \right)$
    - Selection of λ and p
  - Cross validation
    - Should be repeated
  - Permutation test
    - Should include selection of regularization parameter

Non-trivial when more genes than subjects

- → Classification by logistic regression
- → Matlab R2010b
- → Pre-processing of data
  - Concentrations are log-transformed
  - Continuous variables are centralized to zero mean and scaled to one standard deviation
  - Binary variables defined to {-1,1}.
  - Missing data imputed with mean or ML estimate.
- → LASSO regularization
  - Regularization parameter based on cross validation
- → Significance based on permutation test
- → Predictive performance calculated as area under the ROC curve.
  - ROC curves calculated based on double cross validation, regularization in a inner CV loop.

Example:
Classifying gender based on mRNA

- Classify subjects as male/female based on gene expression profile solely.
- For each subject there is 29 gene expression levels,

$$x_{\mathrm{ADA}}, \ldots, x_{\mathrm{VMAT2}}$$

- Predictive probability of gender based on logistic model,

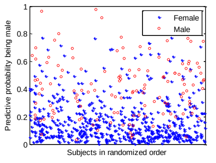
$$\begin{aligned} &\mathrm{P}(\text{Subject } i \text{ is male} \mid \mathrm{X}_{i,\mathrm{ADA}}, \ldots, \mathrm{X}_{i,\mathrm{VMAT2}}) \\ &= 1 - \mathrm{P}(\text{Subject } i \text{ is female} \mid \mathrm{X}_{i,\mathrm{ADA}}, \ldots, \mathrm{X}_{i,\mathrm{VMAT2}}) \\ &= \frac{\exp(\beta_{\mathrm{Const}} + \beta_{\mathrm{ADA}}\mathrm{X}_{i,\mathrm{ADA}} + \beta_{\mathrm{VMAT2}}\mathrm{X}_{i,\mathrm{VMAT2}})}{1 + \exp(\beta_{\mathrm{Const}} + \beta_{\mathrm{ADA}}\mathrm{X}_{i,\mathrm{ADA}} + \beta_{\mathrm{VMAT2}}\mathrm{X}_{i,\mathrm{VMAT2}})} \end{aligned}$$

- Model parameters

$$\beta_{\mathrm{Const}}, \beta_{\mathrm{ADA}}, \ldots, \beta_{\mathrm{VMAT2}}$$

Example:
Estimated model

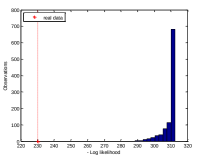
- Model parameters
  - ML estimate
  - LASSO regularization
  - Cross validated regularization parameter



β_{Const} = -1.6654
β_{IL1B} = -0.78074
β_{DPP4} = -0.6967
β_{ARRB2} = 0.58201
β_{CD8B} = -0.57192
β_{CD8A} = 0.45142
β_{GNAS} = -0.42367
β_{IL8} = -0.35866
β_{ARRB1} = 0.3586
β_{SC6A4} = 0.302
β_{MCR} = -0.29382
β_{P2RX7} = 0.25232
β_{MK03} = -0.25176
β_{I23O1} = -0.24763
β_{MK01} = 0.20746
β_{VMAT2} = 0.20369
β_{RGS2} = -0.18348
β_{MK08} = 0.1806
β_{S10AA} = -0.12373
β_{DUS1} = 0.12321
β_{ATF2} = -0.091968
β_{PPCE} = -0.085183
β_{CREB1} = -0.0036507
β_{TSPOA} = 0
β_{DCOR} = 0
β_{MK14} = 0
β_{GCR} = 0
β_{GNAI2} = 0
β_{ADA} = 0
β_{IL6} = 0

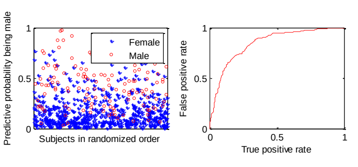
Example:
Significance

- Probability that a model would describe data equally well by chance.
- Permutation test, repeated 1000 times.
- Estimated p-value = 0/1000
- Classifier include 22 genes



Example:
Performance

- Trade-off between sensitivity and specificity
- ROC curve AUC classic measure of predictive power
- AUC = 0.84 for final model on training data
- Cross validated AUC = 0.79 (10-fold repeated 10 times)



Why so complicated?

- Many genes, few subjects
- No clear signal

Method	Comment
Full sample lasso estimation	Too optimistic (performance bias)
CV	To choose smoothing parameter (α)
Repeated CV	To reduce variability in estimation due to random split
Double (outer) CV	To remove bias in performance evaluation
Repeated double CV	To reduce variance in performance evaluation
Permutation test	To give p-value for effect

Summary and Conclusions

- Data with many samples and few subjects
- Different computer intensive techniques in use
  - Simple models
    - Linear regression
    - Logistic regression
  - Regularization
    - LASSO
    - Ridge
    - L0
  - Cross validation
    - Repeated
    - Double cross validation
  - Permutation test

References

- MAQC Consortium (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, Vol. 28, No. 8, pages 827-841.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). The elements of statistical learning, New York, Springer, 2.ed.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.