MedBioInfo - Machine Learning

Statistical principles for supervised learning: Overfitting, regularization and all that - Lab

Computer Lab (with R): A Cancer Modeling Example

Manuela Zucknick

Department of Biostatistics, University of Oslo

manuela.zucknick@medisin.uio.no

May 03, 2022

The example is from Aure et al. (2015), with permission from the authors.

The purpose of modeling

Mathematical models:

- represent an extension of mental models
- often clarify ideas
- may serve as useful simplifications
- are the basis for statistical inference
- encourage comparison with other models
- invite extensions/generalizations

Statistical models:

- are extensions of mathematical models
- also factor in the contribution of unspecified variables

Occam's razor

- Principle from philosophy.
- If two explanations of a phenomenon exists, the simpler one is usually better.
- The simpler model hinges on fewer assumptions, is usually easier to understand, and is easier to falsify
- Always ask yourself if a simple model will suffice, before you decide to move on to a complex method

The linear model

Purpose: study the association between a continuous response y and one or more covariates $x_1, x_2, \ldots, x_m$:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_m x_m + \varepsilon$$

Input data for ith individual:

$$x_{i1}, \ldots, x_{im} \text{ and } y_i$$

Output:

- Estimates $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$
- Standard errors $s_0, s_1, \ldots, s_m$
- P-values for estimates $p_0, p_1, \ldots, p_m$
- P-value for whole model

Machine learning models

- Machine learning methods usually build on a specific model
- Machine learning methods often utilize regularization
- Sometimes, the model is not explicitly given
- Don't be fooled by statements such as "this method does not make any assumptions about the data" sometimes seen e.g. in neural network literature

All models are wrong

Essentially, all models are wrong, but some are useful.

Box, G. E. P.; Draper, N. R. (1987), Empirical Model-Building and Response Surfaces, John Wiley & Sons.

A model is a simplified representation

Just as there is no flat map that is a good representation of the earth's entire surface, there is no single theory that is a good representation of observations in all situations.

Stephen Hawking and Leonard Mlodinow, The Grand Design

Example: assess miRNA-protein expression association

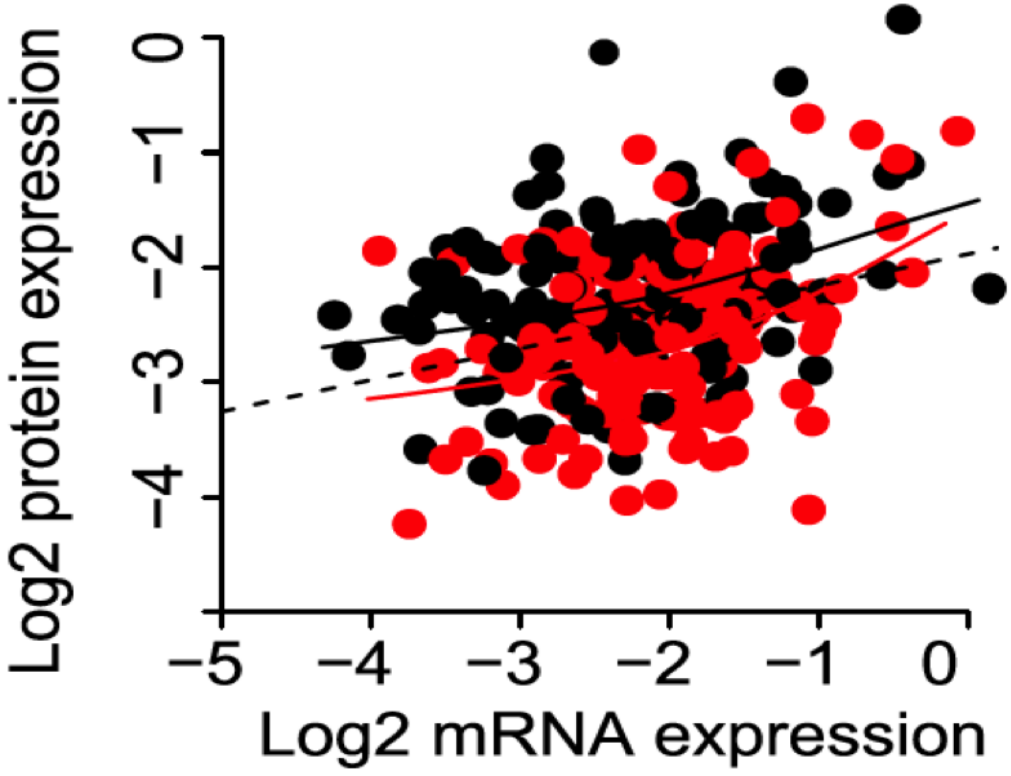
miRNAs

- A microRNA (miRNA) is a small non-coding RNA molecule containing about 22 nucleotides
- Over 2000 miRNAs are known to exist in humans
- miRNAs are known to regulate mRNA expression and protein expression of many genes
- Deregulation of miRNAs have been implicated in cancer initiation and progression

In a recent study we measured the following in $n = 283$ tumors:

- expression of all miRNAs
- expression of all mRNAs
- expression of 105 proteins linked to cancer

Example:



Aure *et al. Genome Medicine* (2015) 7:21
DOI 10.1186/s13073-015-0135-5

Genome Medicine

RESEARCH Open Access

Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer

Miriam Ragle Aure[1,2†], Sandra Jernström[1,2†], Marit Krohn[1,2], Hans Kristian Moen Vollan[1,2,3], Eldri U Due[1,2], Einar Rødland[4,5,6], Rolf Kåresen[7], Oslo Breast Cancer Research Consortium (OSBREAC), Prahlad Ram[8], Yiling Lu[8], Gordon B Mills[8], Kristine Kleivi Sahlberg[2,9], Anne-Lise Børresen-Dale[1,2], Ole Christian Lingjærde[2,5,6*] and Vessela N Kristensen[1,2,10*]

Question

How to assess the relationship between the 105 proteins and the miRNAs?

Consider a specific gene G.
At any given time, a proportion of mRNA is translated into protein and a proportion of protein degrades. Mathematically, this can be modeled as:

$$P'(t) = a \cdot E(t) - b \cdot P(t)$$

where

- $E(t)$ is mRNA expression at time t
- $P(t)$ is protein expression at time t
- a and b are protein-specific rates of translation/degradation

Over a short time interval we may assume $E(t) \equiv E$, and the above equation then has the solution

$$P(t) = (a/b)E + C\exp(-bt)$$

Conclusion: protein expression is proportional to mRNA expression for large t, i.e. $P = (a/b)E$.

Thus a reasonable and flexible model for the mRNA-protein relationship is

$$P = C \cdot E^{\gamma}$$

However, the constant C (the speed of mRNA-protein translation) may depend on miRNAs. We can model this as follows:

$$P = M_1^{\beta_1} \cdot M_2^{\beta_2} \cdots M_k^{\beta_k} \cdot E^{\gamma}$$

where $M_1, \ldots, M_k$ denote the expression of the miRNAs.

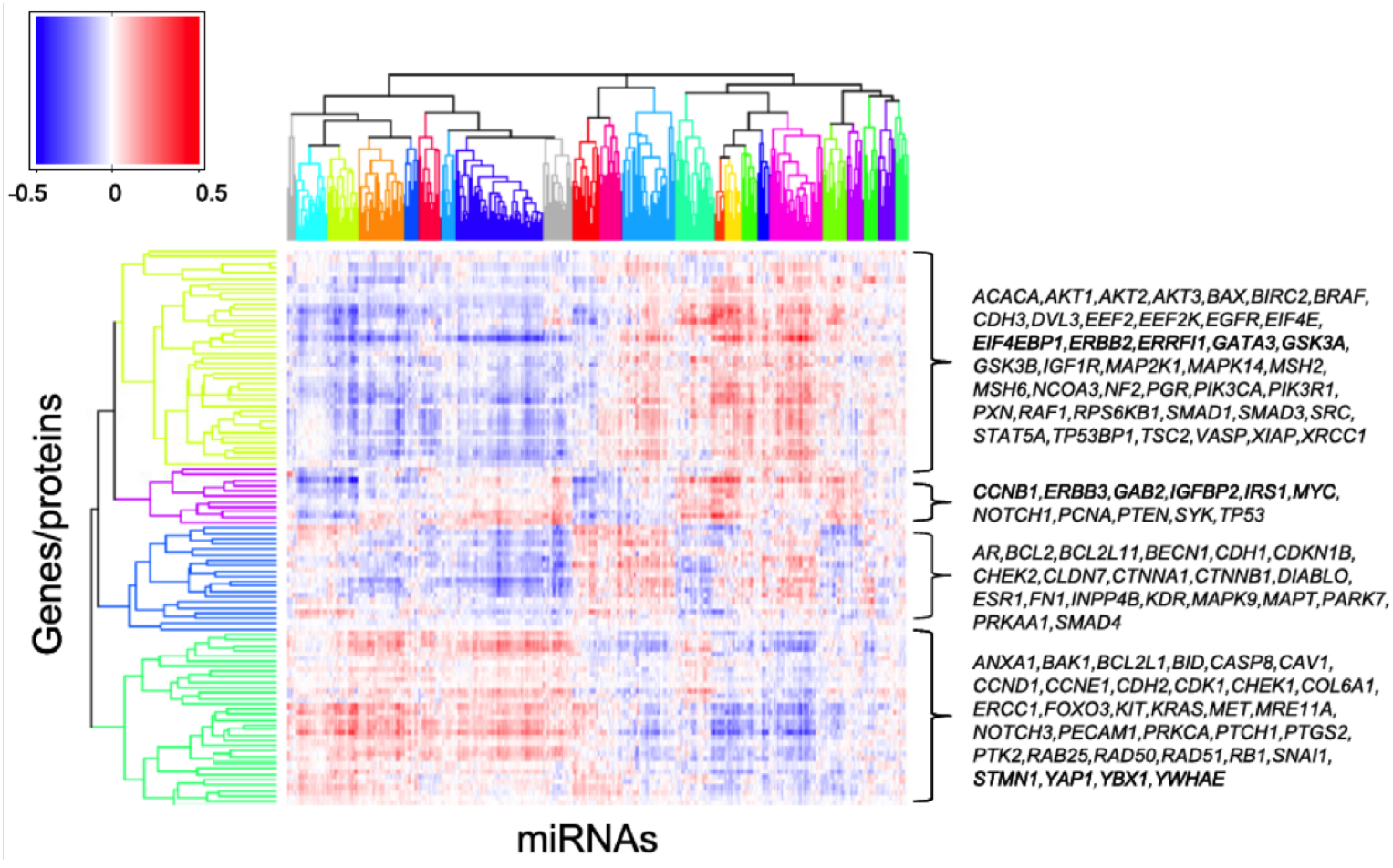
The coefficients $\beta_1, \ldots, \beta_k$ reflect the strengths of the associations between miRNA expressions and protein expression.

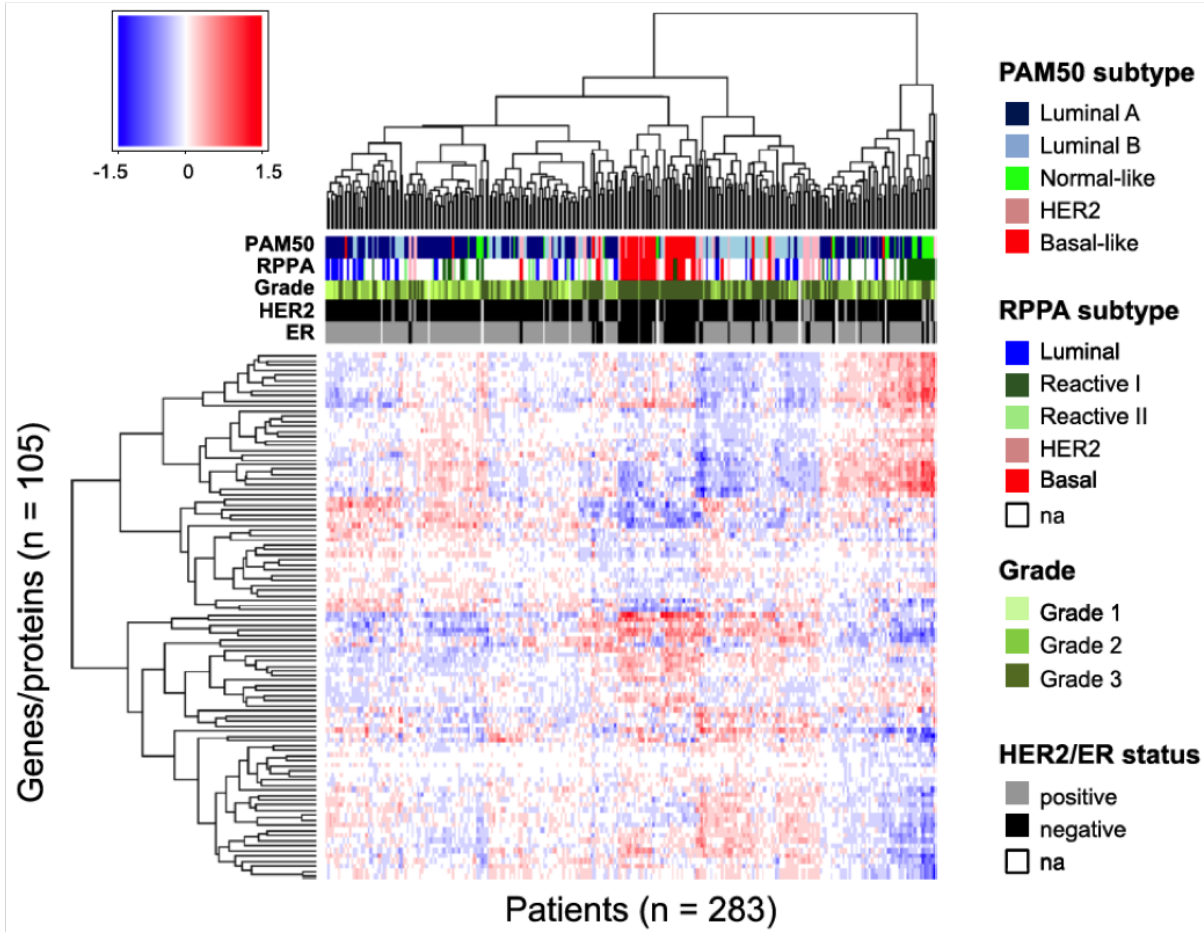
The protein-mRNA-miRNA equation is easier to fit after log-transformation:

$$\log P = \beta_1 \log M_1 + \cdots + \beta_k \log M_k + \gamma \log E + \varepsilon$$

This is a linear model in the log-transformed variables and can be fitted in the usual way.

If the number of variables (miRNAs) is large in the regression equation, we may have to apply some form of regularization, such as ridge regression or the Lasso.


-0.5
0
0.5
Genes/proteins
miRNAs
ACACA,AKT1,AKT2,AKT3,BAX,BIRC2,BRAF,
CDH3,DVL3,EEF2,EEF2K,EGFR,EIF4E,
EIF4EBP1,ERBB2,ERRFI1,GATA3,GSK3A,
GSK3B,IGF1R,MAP2K1,MAPK14,MSH2,
MSH6,NCOA3,NF2,PGR,PIK3CA,PIK3R1,
PXN,RAF1,RPS6KB1,SMAD1,SMAD3,SRC,
STAT5A,TP53BP1,TSC2,VASP,XIAP,XRCC1
CCNB1,ERBB3,GAB2,IGFBP2,IRS1,MYC,
NOTCH1,PCNA,PTEN,SYK,TP53
AR,BCL2,BCL2L11,BECN1,CDH1,CDKN1B,
CHEK2,CLDN7,CTNNA1,CTNNB1,DIABLO,
ESR1,FN1,INPP4B,KDR,MAPK9,MAPT,PARK7,
PRKAA1,SMAD4
ANXA1,BAK1,BCL2L1,BID,CASP8,CAV1,
CCND1,CCNE1,CDH2,CDK1,CHEK1,COL6A1,
ERCC1,FOXO3,KIT,KRAS,MET,MRE11A,
NOTCH3,PECAM1,PRKCA,PTCH1,PTGS2,
PTK2,RAB25,RAD50,RAD51,RB1,SNAI1,
STMN1,YAP1,YBX1,YWHAE


-1.5
0
1.5
PAM50
RPPA
Grade
HER2
ER
Genes/proteins (n = 105)
Patients (n = 283)
PAM50 subtype
Luminal A
Luminal B
Normal-like
HER2
Basal-like
RPPA subtype
Luminal
Reactive I
Reactive II
HER2
Basal
na
Grade
Grade 1
Grade 2
Grade 3
HER2/ER status
positive
negative
na