Detection of Nonlinear Effects in Gene Expression Pathways

Andreas Mayr[1], Djork-Arne Clevert[1,2] and Sepp Hochreiter[1]
[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria
[2]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany

Abstract

Motivation:
Modelling pathways is a central research field in biology. While pathway genes that are mutually linear dependent in their expression values are easy to identify, genes that are nonlinear dependent are hard to find. The detection of such genes is difficult as nonlinearities must be distinguished from noise. We therefore propose an algorithm based on a new developed nonlinear factor analysis algorithm to infer nonlinear gene network components from microarray data.

Results:
We applied our algorithm to the p53 pathway on a number of microarray breast cancer samples and could find some genes that show high nonlinear dependencies across the different datasets.

Goal

Identification of genes depending nonlinearly on the hidden factor

Assumptions

Genes of a pathway are driven by one hidden factor.

Two groups of genes in a pathway:
Linear dependence on hidden factor
Nonlinear dependence on hidden factor

Approach

Nonlinearities: quadratic hidden factor

p-values: probability of a linear gene being detected by chance as nonlinear

Model

Quadratic factor analysis model

$$x = \lambda_0 + \lambda_1 z + \lambda_2 z^2 + \epsilon$$
$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}) \qquad z \sim \mathcal{N}(0, 1)$$

x : Gene expression values
z : Hidden factor (scalar) → pathway activation
ϵ : Independent noise
λ_0 : Mean expression values
λ_1 : Linear gene coefficients
λ_2 : Nonlinear (quadratic) gene coefficients → strength of nonlinearity
Ψ : Diagonal covariance

Model selection

- Model fitting by Expectation-Maximization
- Maximum-a-posteriori solution → priors on linear and nonlinear coefficients
- Higher order moments to identify nonlinearities
- Estimation of moments of the hidden factor:
  - Gaussian approximation
  - New moment-based approximation
  - Importance Sampling
  - Numeric Integration

Availability

- The algorithm is available as a R package.

Experiment: Detection of nonlinear genes in p53 pathway of breast cancer samples

Motivation

- Breast cancer: serious disease affecting a large number of people
- p53 pathway: plays an important role in many types of cancers
- 8 different datasets analyzed

Nonlinear Call

- P-value: linear gene being wrongly detected as nonlinear
- Multiple testing: Bonferroni-adjusted
- Nonlinear call: adjusted p-value threshold of 0.01

Datasets

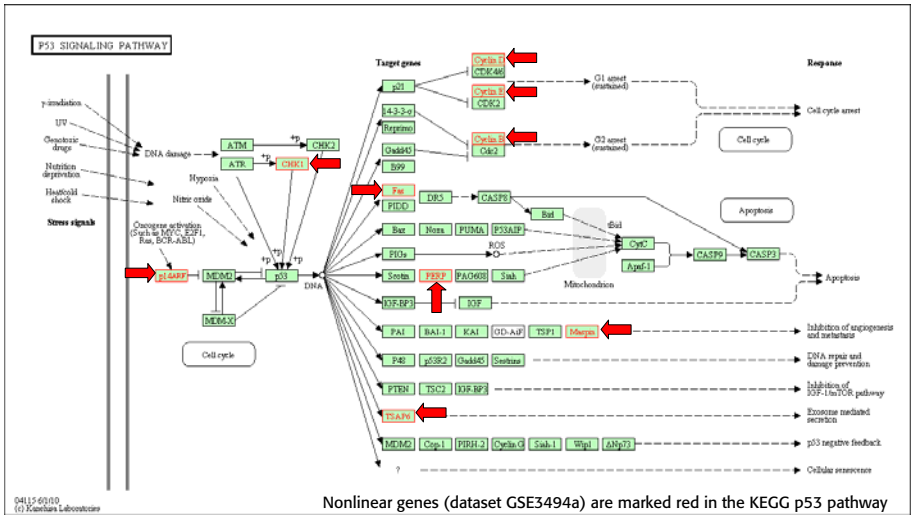
Preprocessing:
- Quantile normalization
- FARMS with use of Brainarray-CDFs
- INI-filtering (to remove uninformative probesets)

Dataset	Samples	Platform	Goal
GSE2109	354	HG-U133_Plus_2	Profiling of tissue samples under standard conditions for the public domain
GSE11121 Schmidt et al. Cancer Res 2008 Jul 1;68(13):5405-13	200	HG-U133A	Search for new prognostic motives in breast cancer
GSE12276 Bos PD et al. Nature 2009 Jan 1;459(7249):1005-9	204	HG-U133_Plus_2	Identification of genes that mediate breast cancer metastasis to the brain
GSE1456a Pawitan Y et al. Breast Cancer Res 2005;7(6):R953-64	159	HG-U133A	Identification of a signature associated with prognosis and impact of adjuvant therapies
GSE1561 Farmer P et al. Oncogene 2005 Jul 7;24(28):4660-71	49	HG-U133A	Prediction of markers for the identification of molecular apocrine breast tumours
GSE3494a Miller al. Proc Natl Acad Sci USA 2005 Sep 20;102(38):13550-5	251	HG-U133A	Development of an expression signature for p53 in breast cancer
GSE4922a Ivshina et al. Cancer Res 2006 Nov 1;66(21):10292-301	289	HG-U133A	Prediction of 264 robust morphological grade-associated markers
GSE6883a Liu et al. N Engl J Med 2007 Jan 18;356(3):217-26	22	HG-U133A	Generation of a 186-gene invasiveness gene signature that is associated with overall survival and metastasis-free survival

Results

Detected nonlinear Genes of the p53 pathway

GSE2109	GSE11121	GSE12276	GSE1456a	GSE1561	GSE3494a	GSE4922a	GSE6883a
CCND1	CCND1	CCND1	CCND1	CCNE1	FAS	FAS	CCND1
CASP3	CCNE1	CASP3	CCNE1	CDKN2A	CCND1	CCND1	CCNG1
CCNE1	CDKN2A	CCNB1	CHEK1		CCNE1	CCNB1	GADD45B
CCNG2	SFN	CCNE1	SERPINB5		CDKN2A	CCNE1	
CDKN2A	IGFBP3	CDC2	THBS1		CHEK1	CHEK1	
CHEK1	SERPINB5	CDK6	CCNE2		SERPINB5	SERPINB5	
GADD45B	STEAP3	CDKN2A	PERP		CCNB2	CCNB2	
SERPINB5	PERP	SERPINB5			STEAP3		
THBS1		RRM2			PERP		
TSC2		CCNB2					
SESN1		CCNE2					
RRM2B		RRM2B					
SHISA5		RFWD2					
		SESN3					



Nonlinear genes (dataset GSE3494a) are marked red in the KEGG p53 pathway

Conclusion

- Significant overlap of nonlinear genes between different datasets
- New insights to biological processes → (re)modelling pathway structures