Endogenous Perturbation Analysis of Cancer - Integrated analysis of mRNA expression and Copy Number Variation.

Rebecka Jörnsten

Mathematical Sciences, University of Gothenburg/Chalmers

jornsten@chalmers.se

November 19, 2010

1. Data Integration - Network modeling
2. Modeling Steady-State mRNA levels
   - Regularization
   - Alternative representation
   - Model Validation
3. Application
   - Analysis of glioblastoma
   - Experimental and Structural Validation
4. Conclusion and Future work

Endogenous Perturbation analysis of Cancer

Goals

- Construct regulatory network and predictive models for cancer pathways
- Identify disease-specific key regulators and their targets

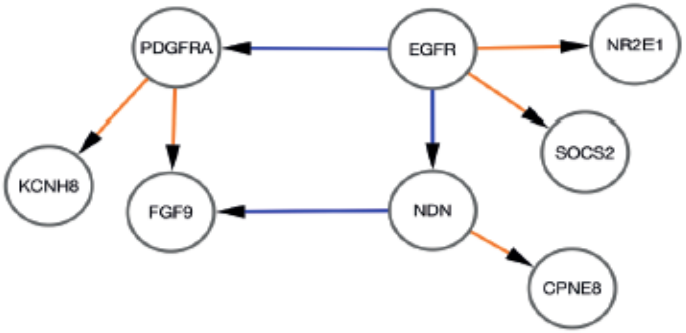
???Is this statistics???

- Challenges in biology: high-throughput technologies (get information on activities of 10000+ genes in different cancer tumors)
- Statistics to help identify which among these 10000+ genes are interacting and how

Endogenous Perturbation analysis of Cancer

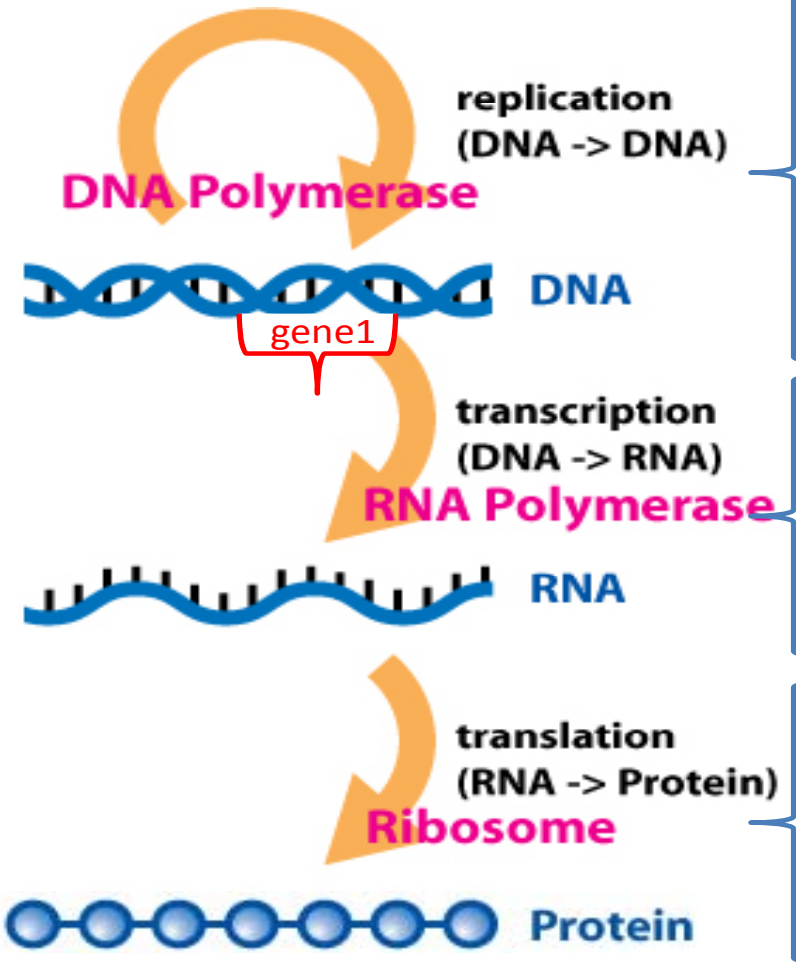
Goals

- Construct regulatory network and predictive models for cancer pathways
- Identify disease-specific key regulators and their targets



- Data now available at multiple levels of these biological processes (genome seq, transcription, proteomics)
- By integrating multiple data sources we hope to identify direct, causal interactions.

Multiple levels of data

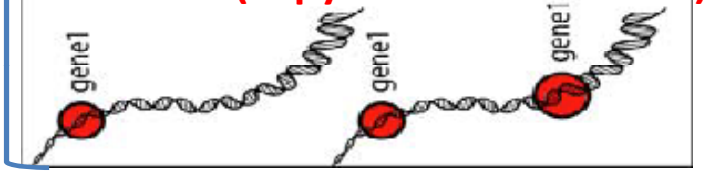


Genetic variation:

* SNPs (change at one locus)

ACGAAA → ATGAAA

* CNVs (copy number variation)



Transcriptional variation:

*measure relative amounts of a gene's mRNA in a cancer tumor compared with a normal cell.

Translational variation:

*measure relative amounts of a gene's protein product

Correlation-based

- Use mRNA expression data only
- Construct networks based on mRNA-mRNA correlation (relationship)
- Pro: massive amounts of mRNA expression data available
- Con: "passive" observation makes network reconstruction difficult

Experimental approaches

- Knock-out of single genes - observe the system
- Pro: targeted perturbation and question
- Con: expensive, limited amounts of data

Using naturally occurring genetic variation

- Genetic variation: in a population of individuals, different yeast strains, crosses

Endogenous perturbations

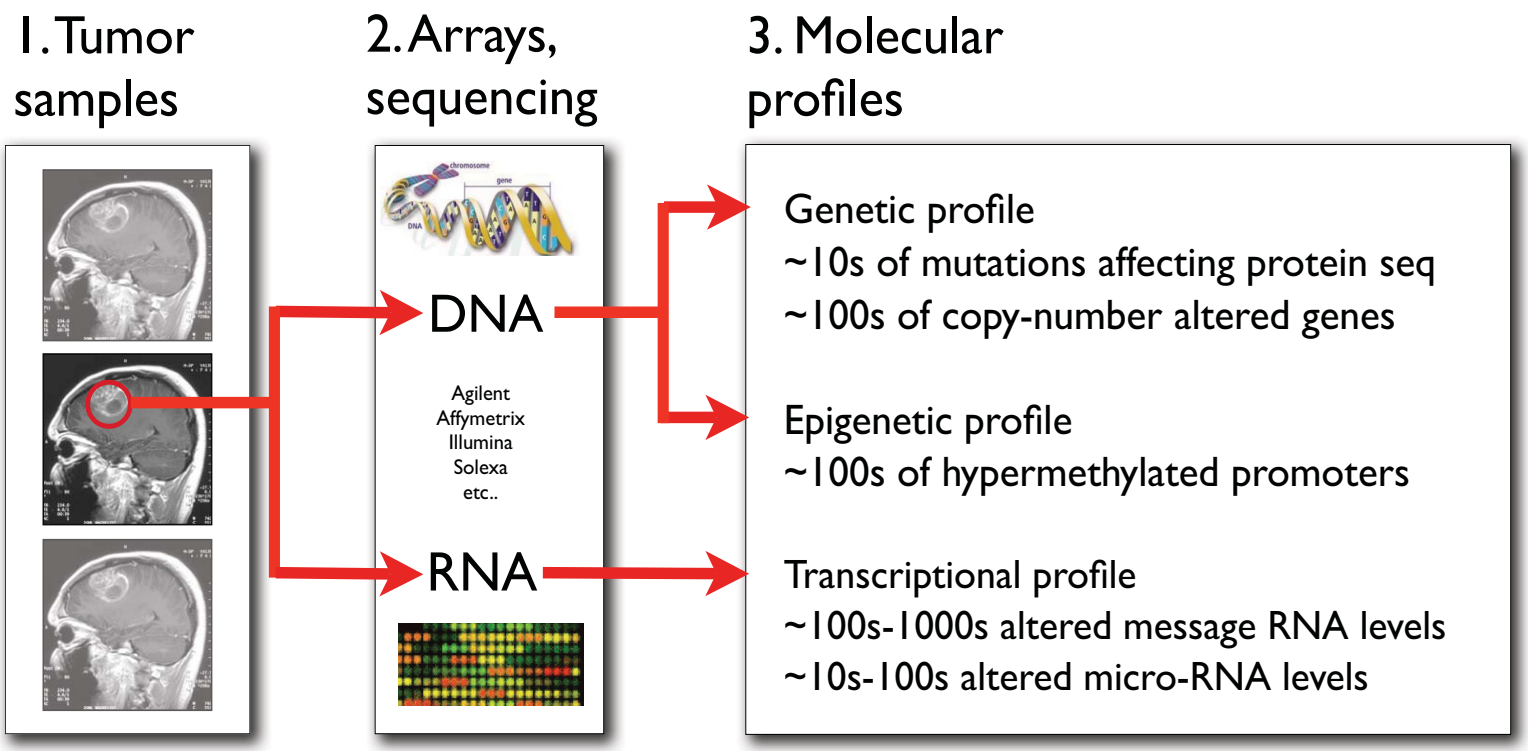
- Focus on *acquired* genetic variation, e.g. mutations in tumors.
- View this as a perturbation of the biological system and observe the response

Endogenous perturbation

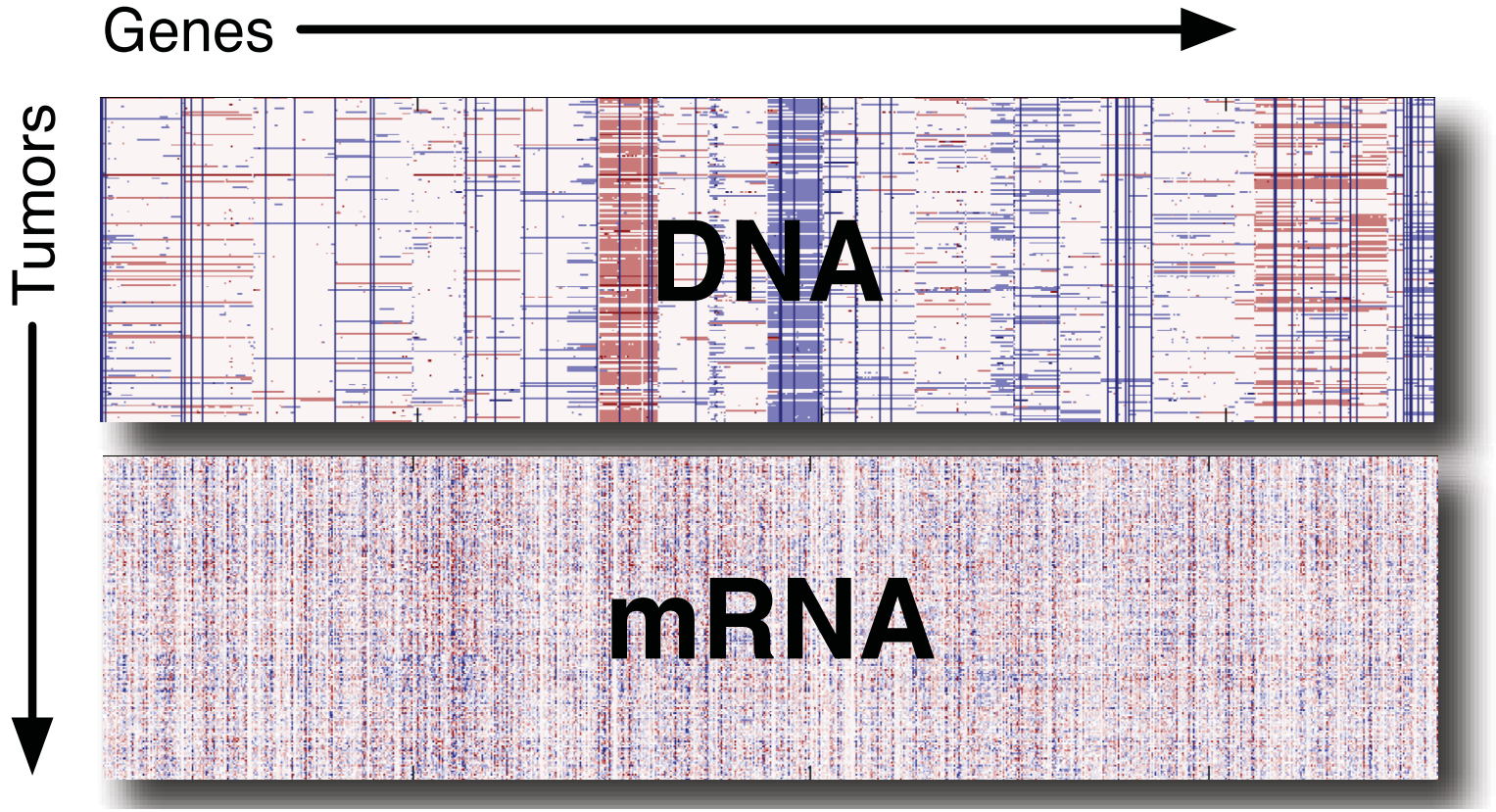
- We view each tumor's Copy Number Aberration (CNA) profile as a system perturbation that simultaneously affects multiple genes, and
- the mRNA profiles as the steady-state response to that perturbation

- CNAs tend to appear in a patient-specific and multifactorial manner - ideal for network construction
- Ongoing projects like The Cancer Genome Atlas means massive amounts of mRNA/CNA data are available

Data - 186 tumors from the Cancer Genome Atlas consortium (TCGA)



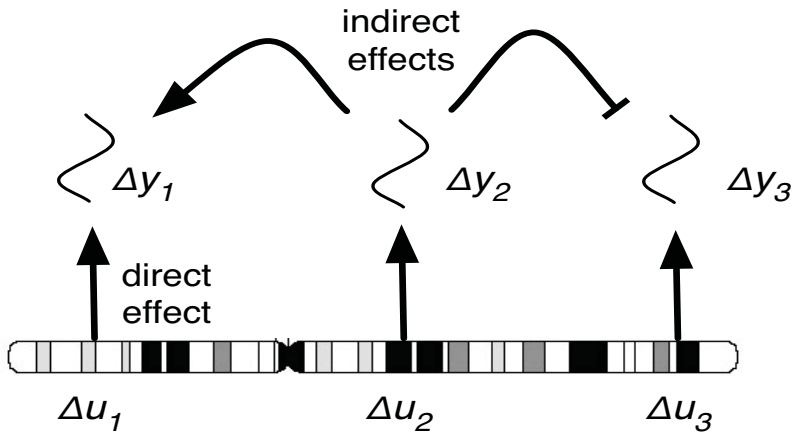
Data - 186 tumors from the Cancer Genome Atlas consortium (TCGA)



Model with multiplicative effects

$$\frac{dy_i}{dt} = u_i \alpha_i \prod_j y_j^{w_{ij}} - \beta_i \prod_j y_j^{v_{ij}}$$

- where y_i is the mRNA expression of gene i, and u_i the copy number.
- Elements w_{ij} and v_{ij} denote the effect of gene j on i during synthesis/degradation respectively,
- and the α_i and β_i denote "environmental effects (non-CNA perturbations)"



Steady-state solution

- We take CNA profile u and a baseline reference $\tilde{u}$, and likewise for mRNA y, $\tilde{y}$.
- Define $\Delta u_i = \log(u_i) - \log(\tilde{u}_i)$, $\Delta y_i = \log(y_i) - \log(\tilde{y}_i)$
- At steady-state we can write:

$$\Delta u_i + \sum_j (w_{ij} - v_{ij}) \Delta y_i +$$

$$+ \overbrace{(\log(\alpha_i) - \log(\tilde{\alpha}_i)) - (\log(\beta_i) - \log(\tilde{\beta}_i))}^{\gamma_i} = 0$$

- We denote the *direct* causal influence of transcript j on i by $a_{ij} = w_{ij} - v_{ij}$.
- We want to find the "important" a_{ij} (i.e. large/real effects)

For each gene we can now write a linear equation that represents this genes mRNA expression:

Steady-state solution

$$a_{ii} y_i = \sum_{j \neq i} -a_{ij} y_j - u_i + \gamma_i$$

Note: we dropped the Δy symbol here - for convenience we just use y to denote the baseline adjusted mRNA and similarly for CNA

Regression modeling

- $a_{ii} y_i = \overbrace{\sum_{j \neq i} -a_{ij} y_j - u_i}^{model} + \gamma_i$ (from now on, set $a_{ii} = 1$)
- We want to use data from the 186 tumors to *estimate* the a_{ij}

Data

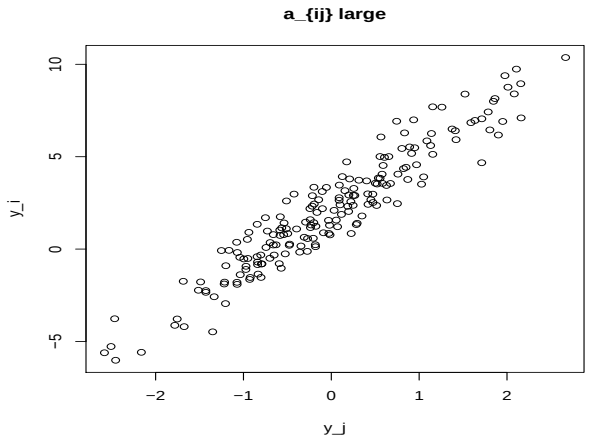
- Observe mRNA and CNA for tumors $t = 1, \cdots, T = 186$ and genes $i = 1, \cdots, n = 10000+$.
- We assume the same a_{ij} describe the mRNA expression for all tumors and
- the term γ_i captures the tumor specific (unmeasured) variation = referred to as "noise" or "error".
- Data: y_{it}, u_{it}.
- Find the a_{ij}s that best match up the ys and the us.

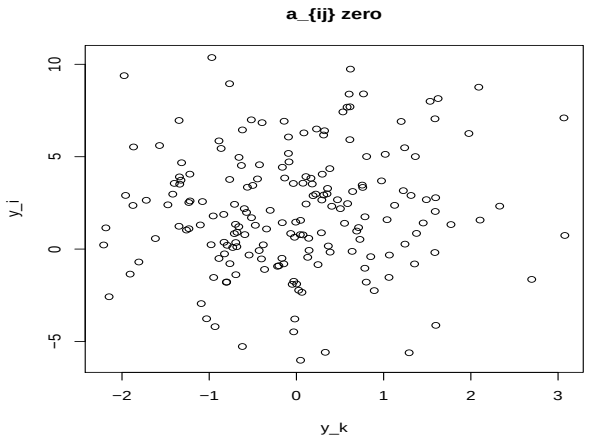
Regression modeling

- Least Squares: Find a_{ij}s to minimize

$$\sum_{t=1}^{T}(y_{it} - (\sum_{j \neq i} -a_{ij}y_{jt} - u_{it}))^2 = \sum_{t=1}^{T}\gamma_{it}^2$$

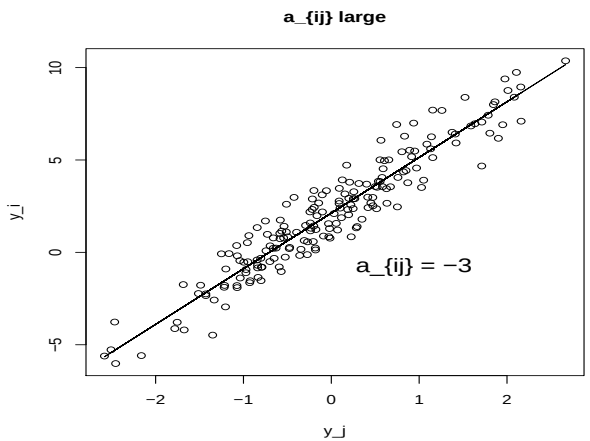
- That is: we minimize the average "noise" or errors over all tumors, i.e. how the observed mRNA y_i deviates from the model a_{ij}

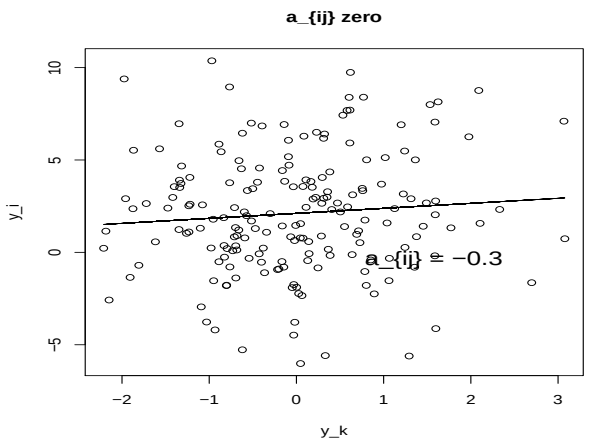




Regression modeling

- When we minimize the squared errors we get an *estimate* of the *true* a_{ij}
- Which interactions are real ($a_{ij} > 0$ or < 0), and which are not ($a_{ij} = 0$)?
- There are statistical tests for deciding if an estimated a_{ij} is large enough to be declared "real" (depends on the magnitude of the estimated a_{ij}, the number of samples(tumors), the amount of noise, ...).





Multivariate response regression modeling

- For the whole system we can write

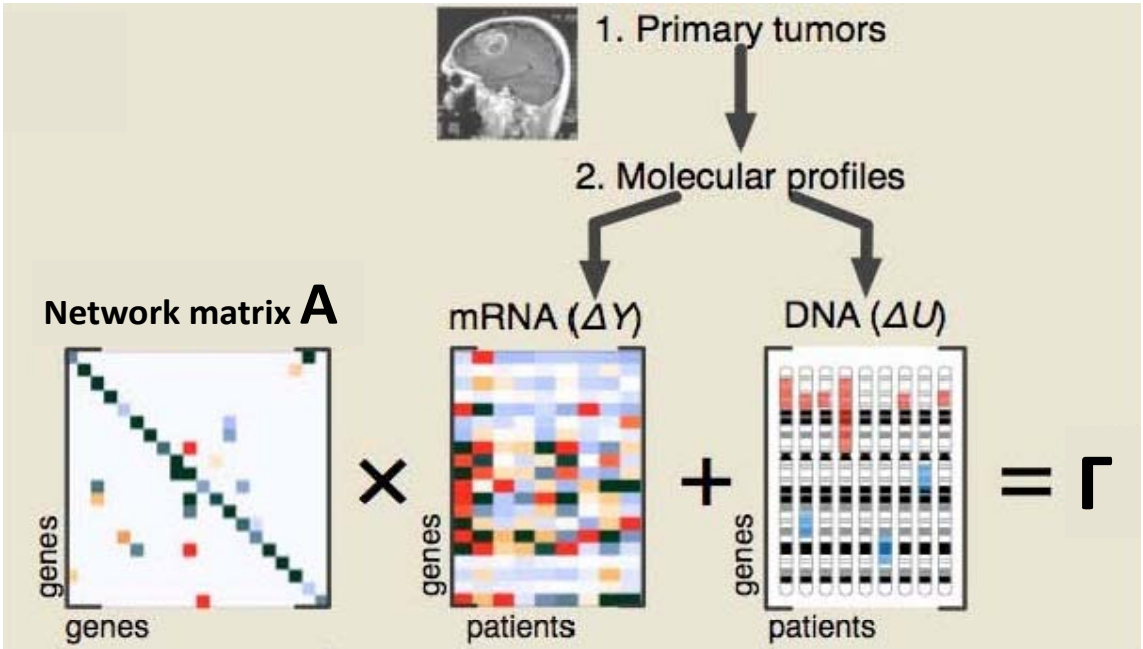
$$Ay + u = \Gamma$$

- That is, row i in A contains the interactions that gene i has with the other genes a_{ij}, etc.
- We are solving 10000+ regression problems!!!

Network solution

$$A\mathbf{y} + \mathbf{u} = \Gamma$$

- A is called the *network matrix*, our parameters of interest.
- Γ includes all the tumor-specific differences, unmeasured environmental effects, SNPs etc.



Oh-oh! Another complication

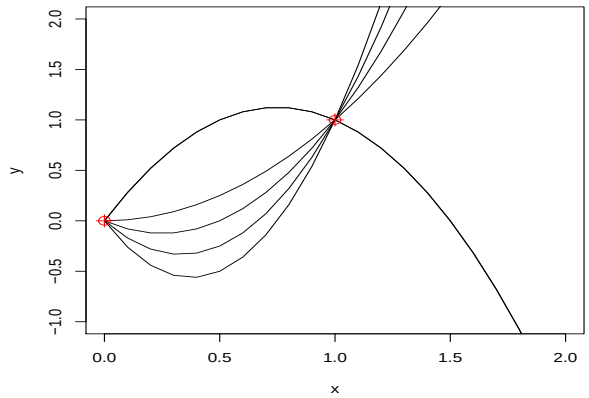
Let's go back to the original regression model for gene i

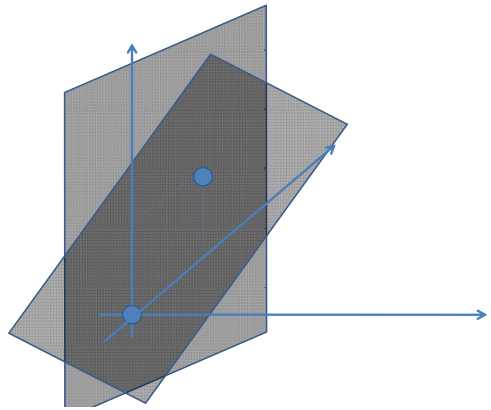
$$a_{ii} y_{it} = \overbrace{\sum_{j \neq i} -a_{ij} y_{jt} - u_{it}}^{model} + \gamma_{it}$$

- We have data from 186 tumors
- We have $p = 10000+$ unknowns in the model (all the a_{ij}'s)
- This is called the $p > T$ problem - we are asking more questions about the data than we have observations.
- Consequence: there is an infinite number of solutions a_{ij}'s that fit the data equally well!!!

$p > T$ problem

Model	Model
$y = \alpha + \beta x + \gamma x^2$	$y = \alpha + \beta x_1 + \gamma x_2$
Data	**Data**
$(x, y) = \{(0,0), (1,1)\}$	$(x1, x2, y) = \{(0,0,0), (0,1,1)\}$
Solution	**Solution**
$\alpha = 0, \gamma = 1 - \beta$	$\alpha = 0, \gamma = 1 - \beta$





More formal

$$y = Xa + \gamma$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ y_T \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & . & x_{1,10000} \\ x_{2,1} & x_{2,2} & . & x_{2,10000} \\ . & . & . & . \\ x_{T,1} & x_{T,2} & . & x_{T,10000} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ . \\ a_{10000} \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ . \\ \gamma_T \end{pmatrix}$$

- Each gene takes its turn to be y, and the other genes form columns in X.
- Note $p = 10000$ genes $> T = 186$ samples

Least Squares

We want to minimize

$$Q(a) = \sum_{t=1}^{T} \gamma_t^2 = (\gamma_1 \gamma_2 . \gamma_T) \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ . \\ \gamma_T \end{pmatrix} = \gamma^T \gamma = (y - Xa)^T (y - Xa)$$

$$\frac{dQ}{da} = -2X^T(y - Xa) = 0 \rightarrow (X^T X)a = X^T y$$

When $p > T$ (here 10000 vs 186) the inverse of $X^T X$ does not exist! No unique solution a

Solution 1 - subset of genes

Pick max T genes to be in the model. Then

$$\hat{a} = (X_S^T X_S)^{-1} X_S^T y$$

where $\hat{a}$ means we have estimated the interactions a, and X_S means we are only using a subset S of the $p = 10000$ columns in X.

- PRO: Simple and interpretable model $\hat{a}$ - "genes in set S affect gene y"
- CON: Which set S to use? How large - max T - but perhaps better to use fewer genes?
  There are 2^{10000} possible combinations of genes we can put in set S!!!

Solution 2 - Regularization

We "fix" the inverse of $X^T X$:

$$\hat{a} = (X^T X + \lambda I)^{-1} X^T y$$

That is, we add something to the diagonal of $X^T X$. Now we can take the inverse!
Note, $\hat{a} \simeq a_{original}/(1+\lambda)$ - That is, we are *shrinking* the estimates of a toward 0.

- PRO: Simple fix to the inverse problem
- CON: We made all the as smaller - real effects and the a's that are truly zero
  What should λ be?

Solution 2 - Regularization

This is also called "regularized regression" because we can arrive at this solution by solving problem:

$$\min_a Q(a) = (y - Xa)^T(y - Xa) + \lambda \sum_{j=1}^{10000} a_j^2 = (y - Xa)^T(y - Xa) + \lambda a^T a$$

That is, we *penalize* large values of a (err on the side of caution).
Solution:

$$\frac{dQ}{da} = -2X^T(y - Xa) + 2\lambda a = 0 \rightarrow (X^T X + \lambda I)a = X^T y$$

We "fix" the inverse of $X^T X$:

$$\hat{a} = (X^T X + \lambda I)^{-1} X^T y$$

Solution 3 - Lasso regression

This idea on penalizing the large a values can be generalized to other types of penalties.
Last 10 years, lots of statistics research into penalizing the *absolute values* of a: "Lasso regression"

$$\min_a Q(a) = (y - Xa)^T(y - Xa) + \lambda \sum_{j=1}^{10000} |a_j|$$

Turns out the solution looks something like this:

- $\hat{a}_j = a_{original} - 1/\lambda$ if $\hat{a} > 1/\lambda$
- $\hat{a}_j = a_{original} + 1/\lambda$ if $\hat{a} < -1/\lambda$
- $\hat{a}_j = 0$ otherwise

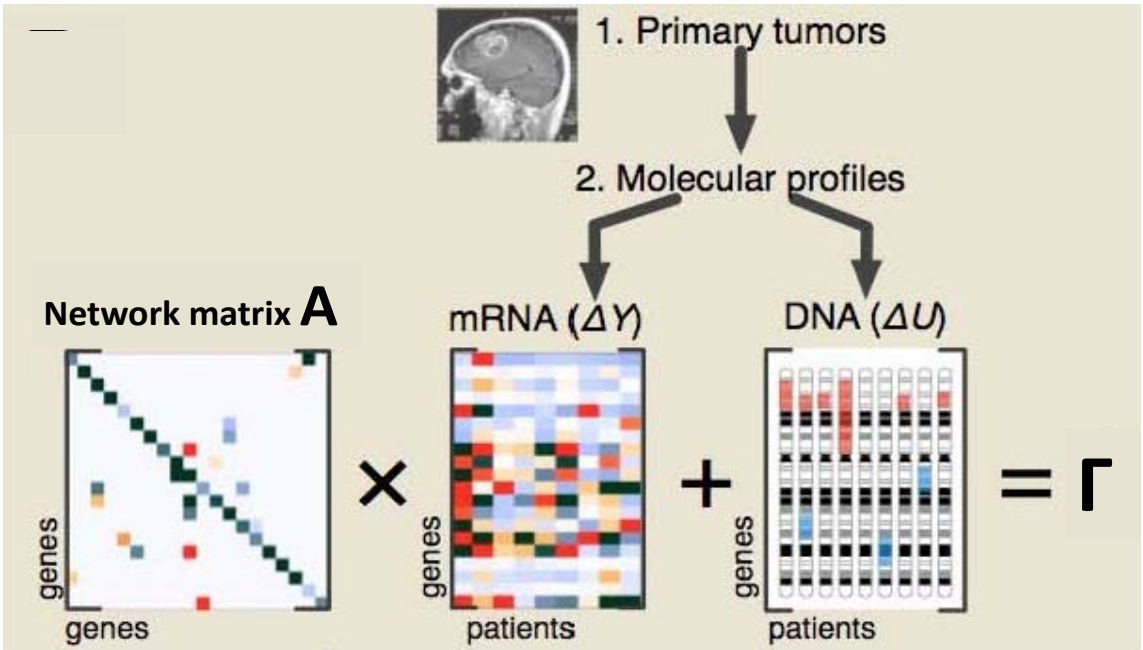
That is, small a's are set to 0, large a's are shrunk by $1/\lambda$

- PRO: Combines the best of both worlds: selection of non-zero a's, and simple procedure involving just λ rather than the 2^{10000} combinatorics
- CON: What should λ be?

Finally! Back to Network solution

$$A\mathbf{y} + \mathbf{u} = \Gamma$$

- We solve for A one row (gene) at a time
- For each gene we solve a Lasso regression problem
- Still need to choose λ - we'll come back to that in a minute.

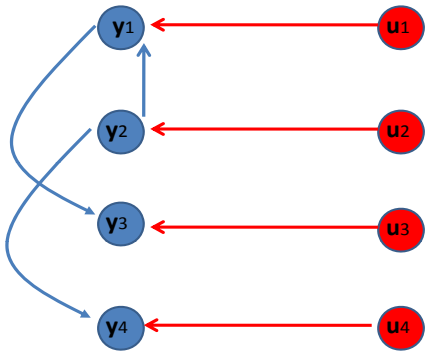


Alternative Representation

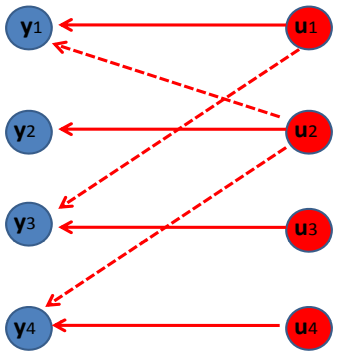
$$\mathbf{y} = G\mathbf{u} + \Gamma'$$

- $G = -A^{-1}$ is called the *system matrix*.
- G represents the *system gain*: where the genetic variation (system input) shows up as amplified (positive or negative) signal in the mRNA expression (system output).

Alternative Representation



$$
\begin{aligned}
y_1 &= a_{12} y_2 + u_1 + \gamma_1 \\
y_2 &= u_2 + \gamma_2 \\
y_3 &= a_{31} y_1 + u_3 + \gamma_3 \\
y_4 &= a_{42} y_2 + u_4 + \gamma_4
\end{aligned}
$$

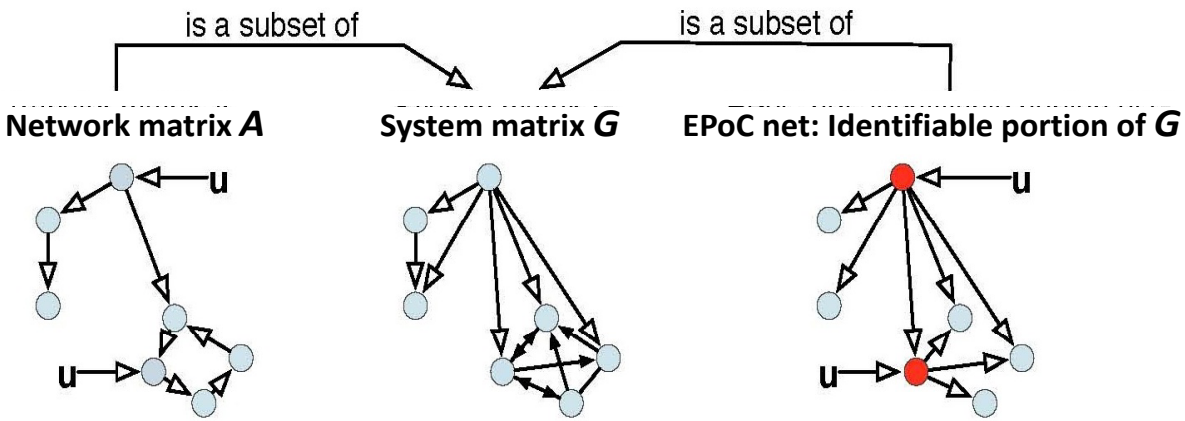


$$
\begin{aligned}
y_1 &= u_1 + \overbrace{a_{12}}^{g_{12}} u_2 + \gamma_1' \\
y_2 &= u_2 + \gamma_2' \\
y_3 &= \overbrace{a_{31}}^{g_{31}} u_1 + \overbrace{a_{31} a_{12}}^{g_{32}} u_2 + u_3 + \gamma_3' \\
y_4 &= \overbrace{a_{42}}^{g_{42}} u_2 + u_4 + \gamma_4'
\end{aligned}
$$

Alternative Representation - Why?

$$A\mathbf{y} + U = \Gamma, \quad \mathbf{y} = G\mathbf{u} + \Gamma'$$

- The genetic variation is thought to be (in part) the *disease cause*, whereas the mRNA expression is the symptom.
- Looking at G is therefore more informative for understanding which part of the biological network is disease specific.
- Practical reason: A is difficult to estimate due to strong correlations (lots of models fit the data equally well).



Choosing λ - the network size

- Remember: λ controls which a's (g's) are zero and which are non-zero.
- Large λ: we risk missing some important interactions in our model
- Small λ: we risk including "false" interactions in our model.
- How do we validate the results?

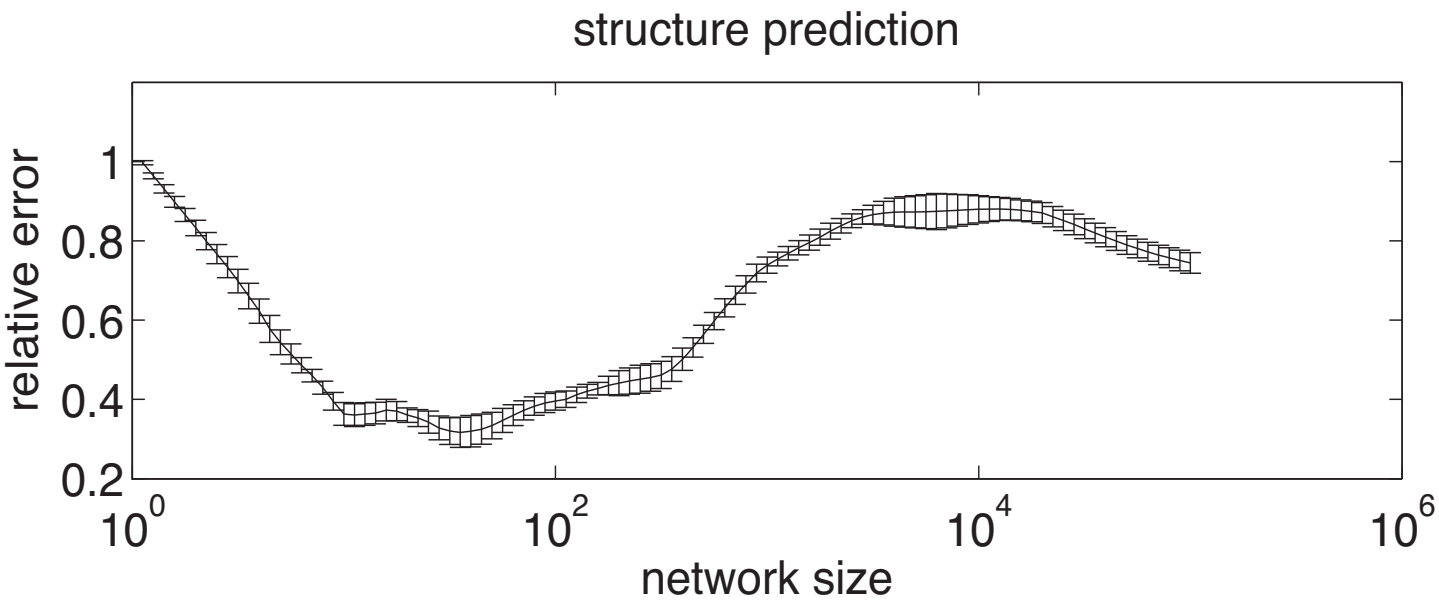
Validation statistics

We consider two different evaluation measures;

- Network structure consistency - Kendall's W
- mRNA prediction - minimum average prediction error

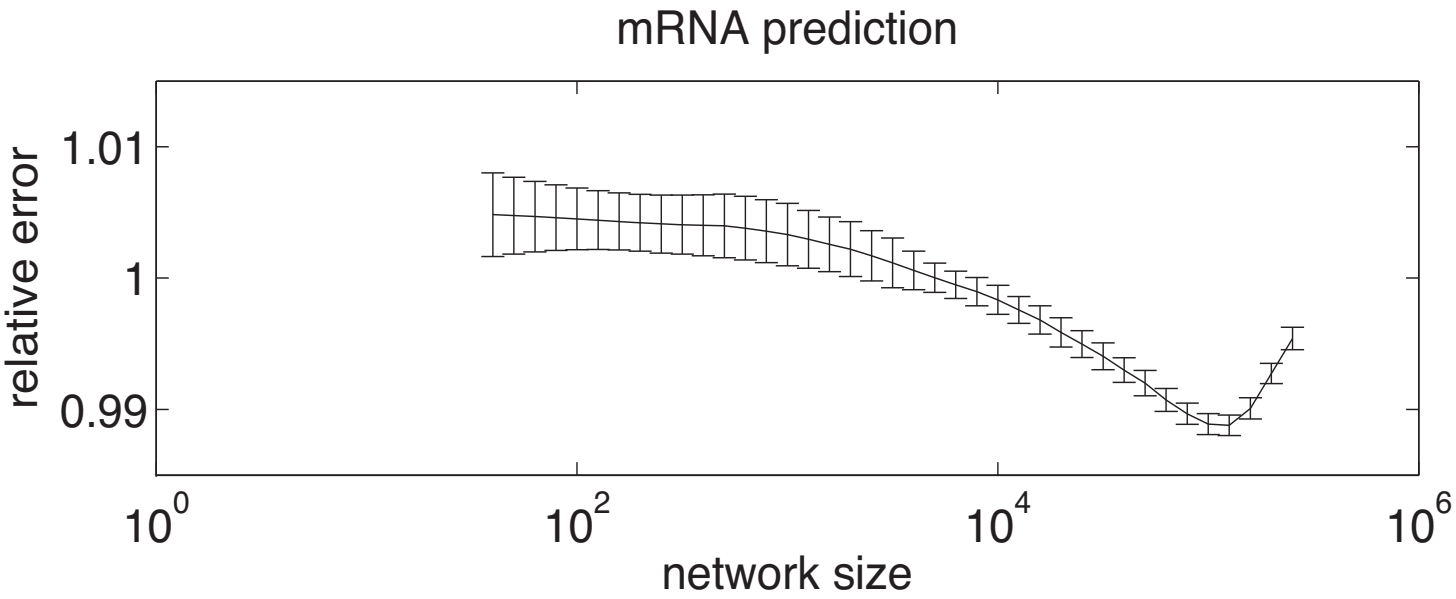
Network consistency

- We randomly split the data into two halves, and estimate the network on each of the two sets of data.
- We compare the network agreement between the two data sets using Kendall's W: measures how well the non-zero and zero a's (or g's) agree in terms of location and magnitude.
- Most consistent networks contain $\sim$ 400 edges.

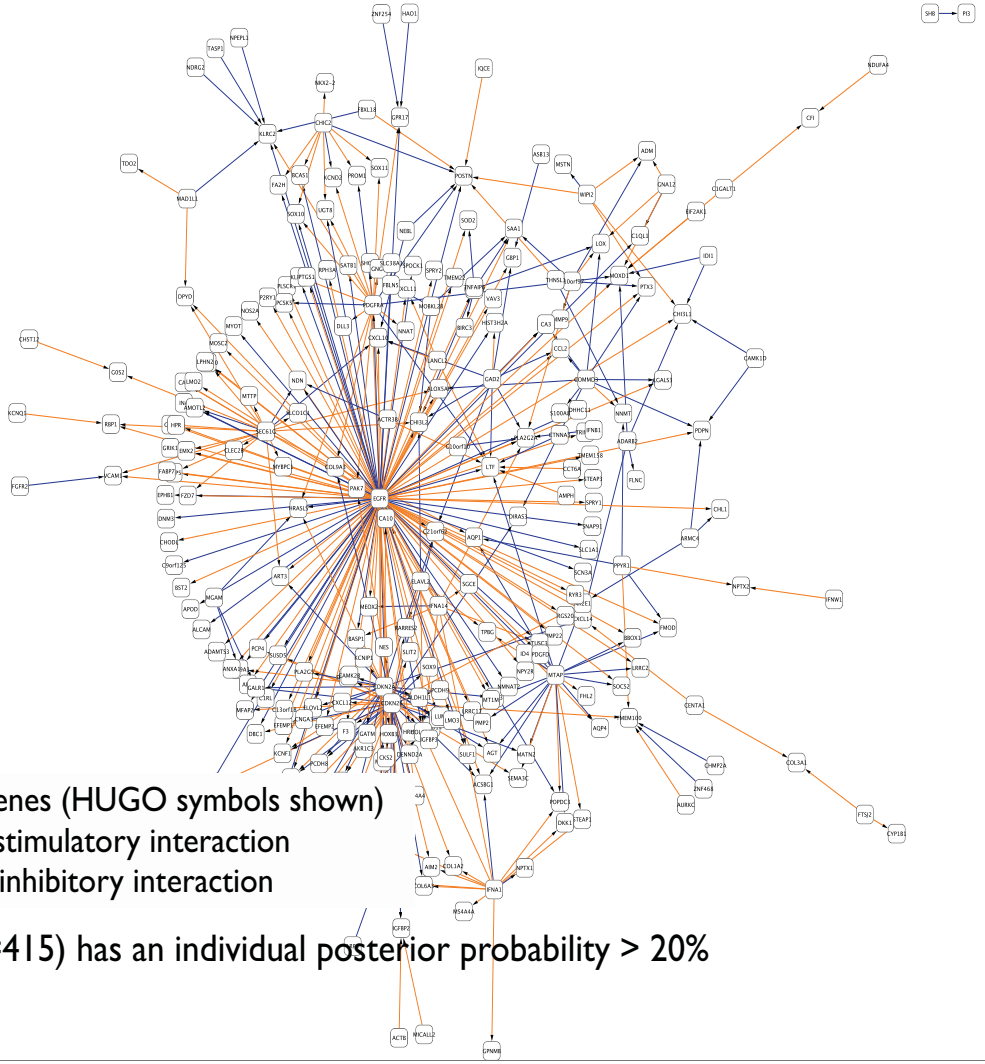


mRNA prediction

- We split the data into random halves, and estimate a network on each of the two sets of data.
- We use the network model from one half to predict the mRNA levels in the other half, and vice versa.
- Minimum mRNA prediction errors for networks with $\sim$ 10000 edges.



The EPoC model of grade IV glioblastoma is based on 146 patients from the Cancer Genome Atlas study

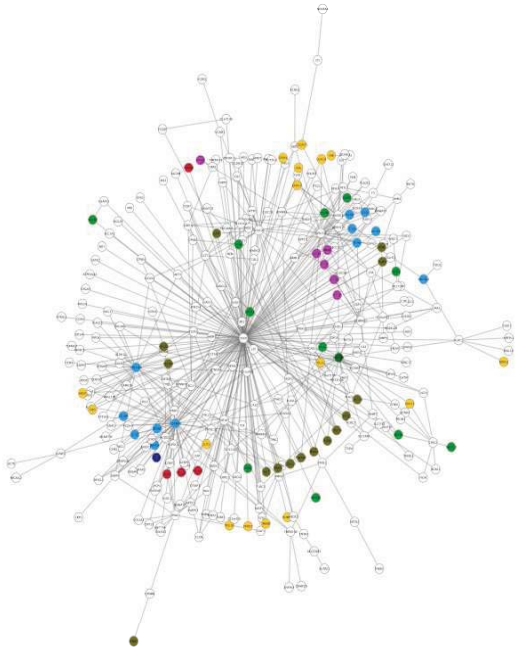


Circles are genes (HUGO symbols shown)
Red arrows: stimulatory interaction
Blue arrows: inhibitory interaction

Each interaction (n=415) has an individual posterior probability > 20%

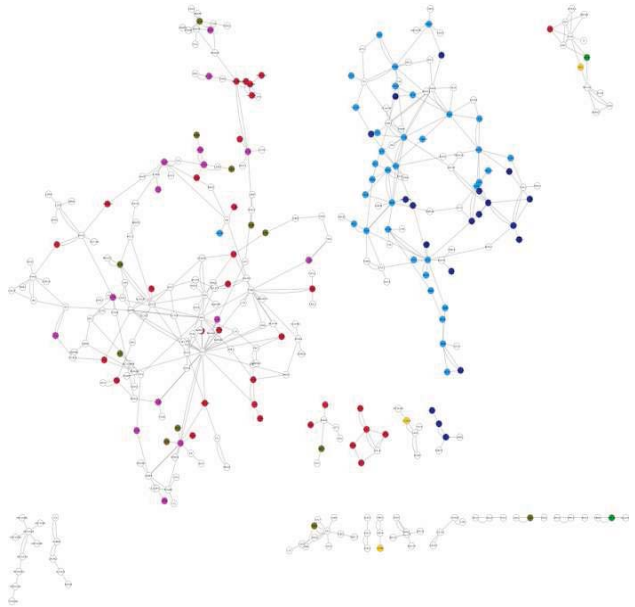
Glioma network analysis

EPoC, *G* matrix solution



- cell differentiation (GO:0030154, 2.29e-09)
- nervous system development (GO:0007399, 1.35e-24)
- cell-cell signaling (GO: GO:0007267, 8.39e-25

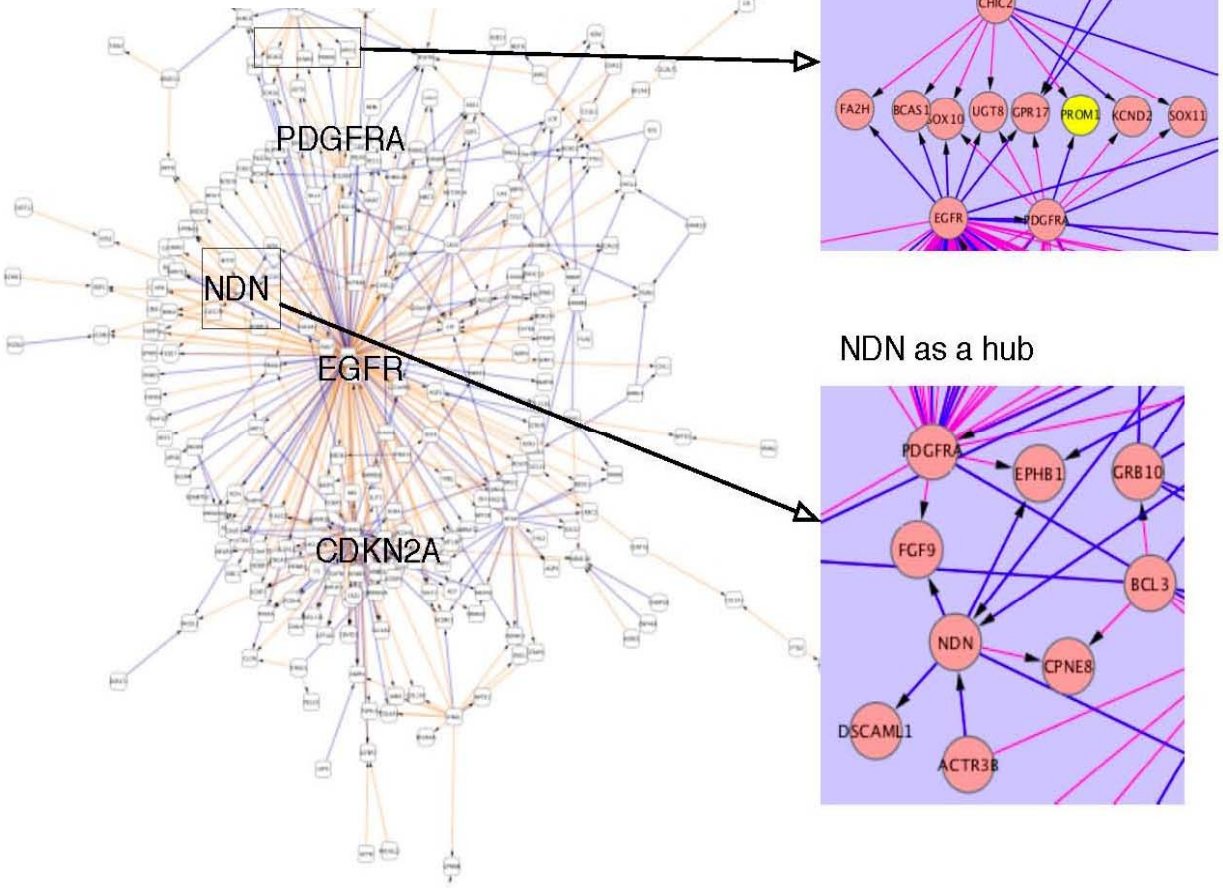
EPoC, *A* matrix solution



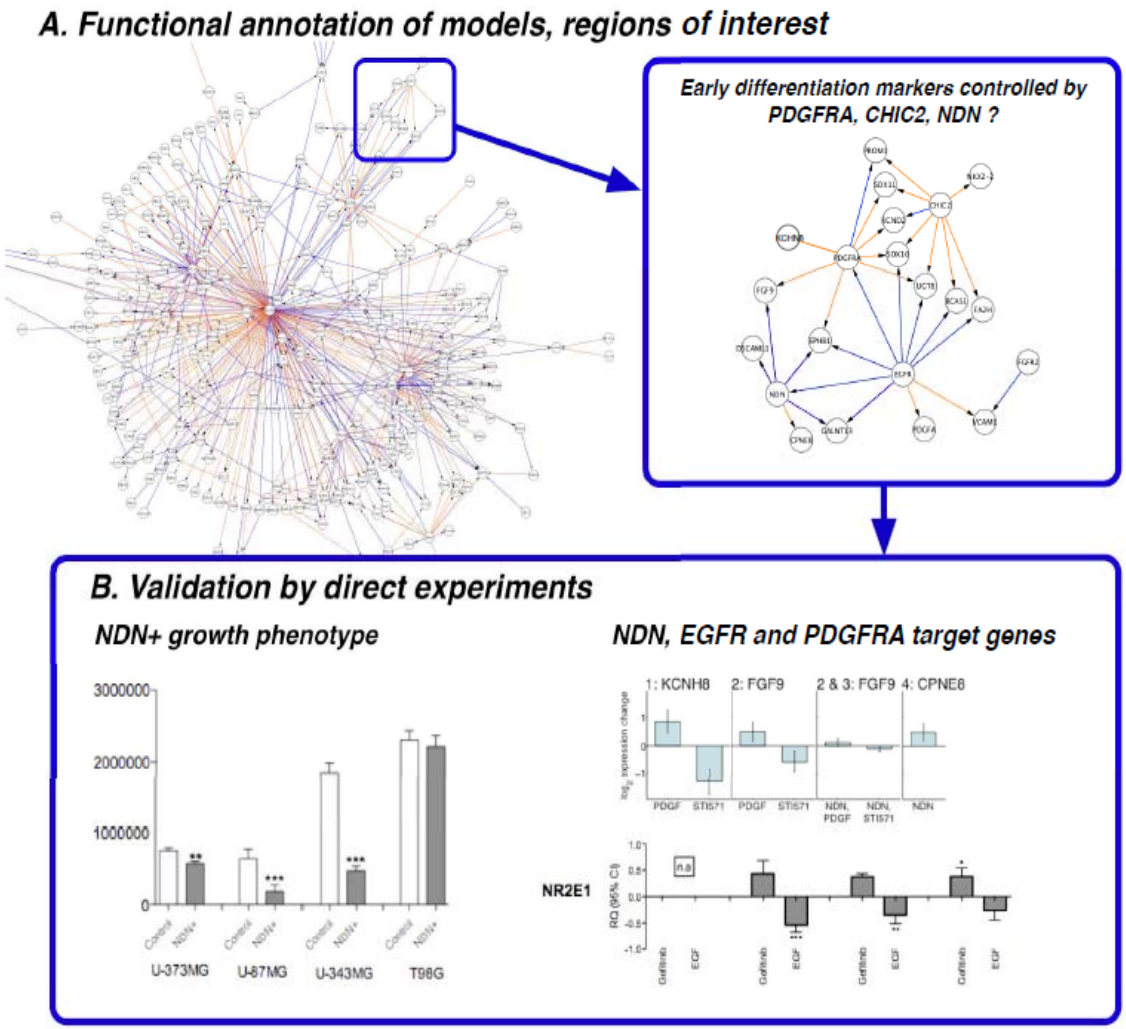
- inflammatory response (GO:0006954, 1.16e-19)
- immune response (GO:0006955, 3.3e-39
- cell cycle (go: GO:0007049, 2.3e-65)

Glioma network analysis

The overall network contains a number of well-established regulators and markers



Experimental validation

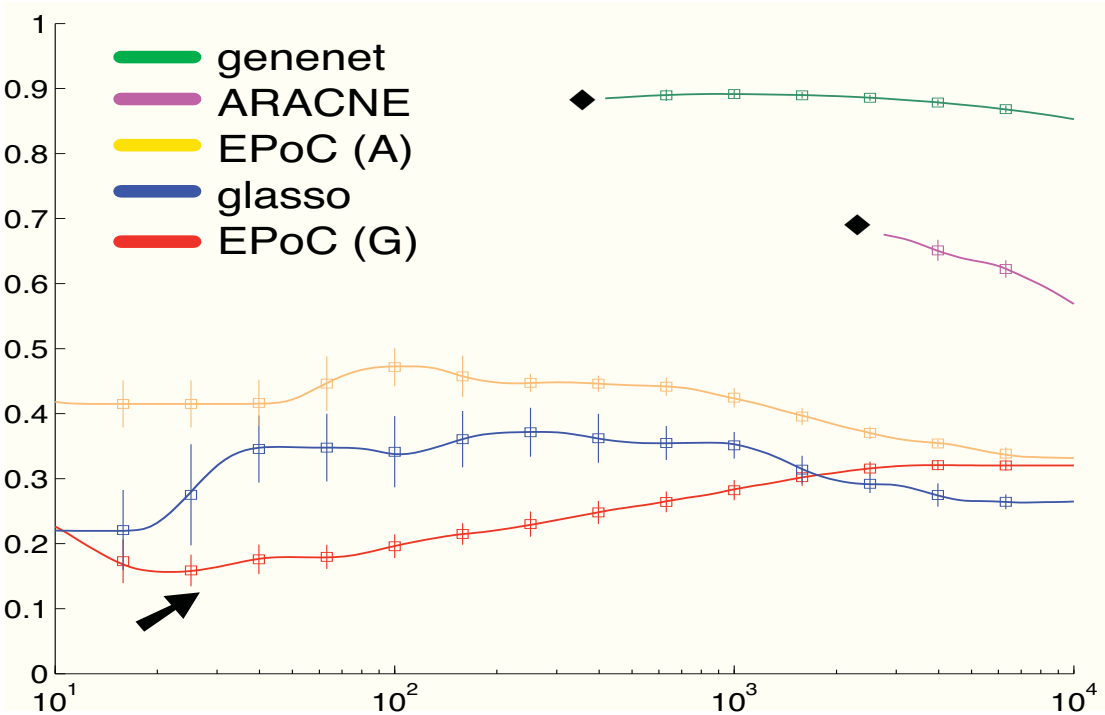


Network consistency

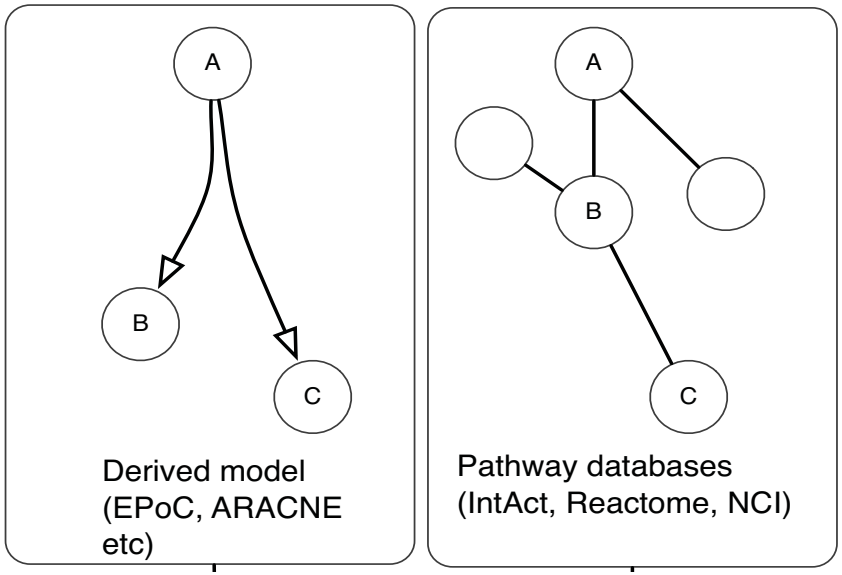
For 146 tumors, we have mRNA and CNA data from

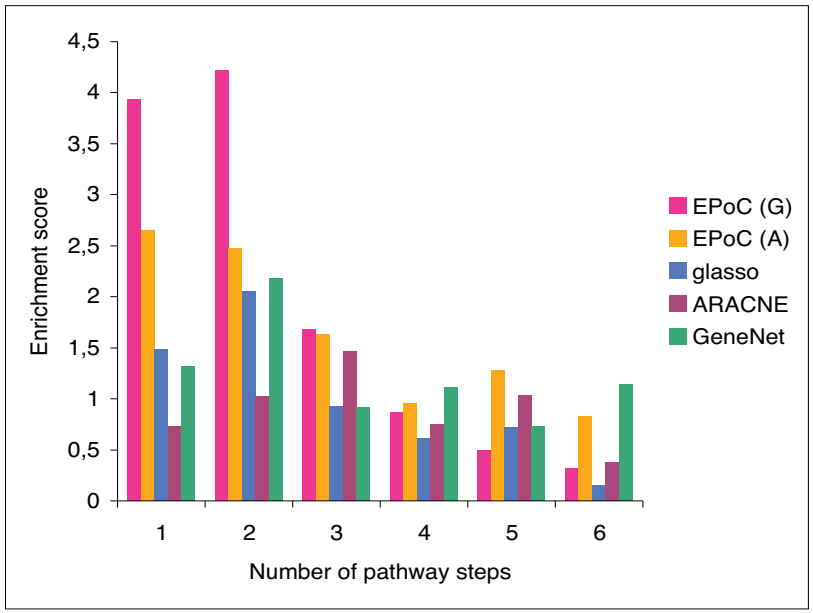
- Two different labs
- Two different platforms (agilent, affy)

We compare network consistency across network sizes and between competing methods.



Overlap with pathway databases



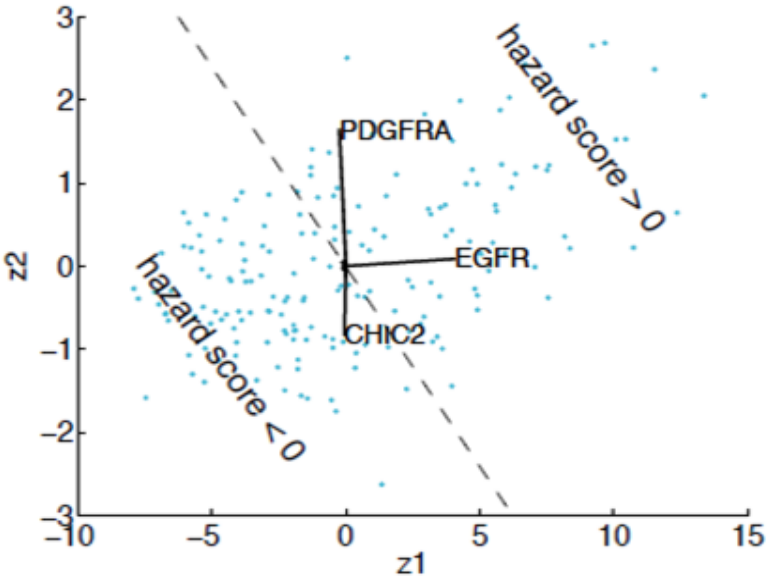


- We map interactions found by different methods to molecular links in pathway repositories HPRD, Reactome, Intact, and NCI-nature.
- Compare pathway links to the shortest paths in networks.
- EPoC-G is clearly enriched for short or direct paths. EPoC-A, GLasso, ARACNE and GeneNet are associated with longer paths.

Decomposition of G

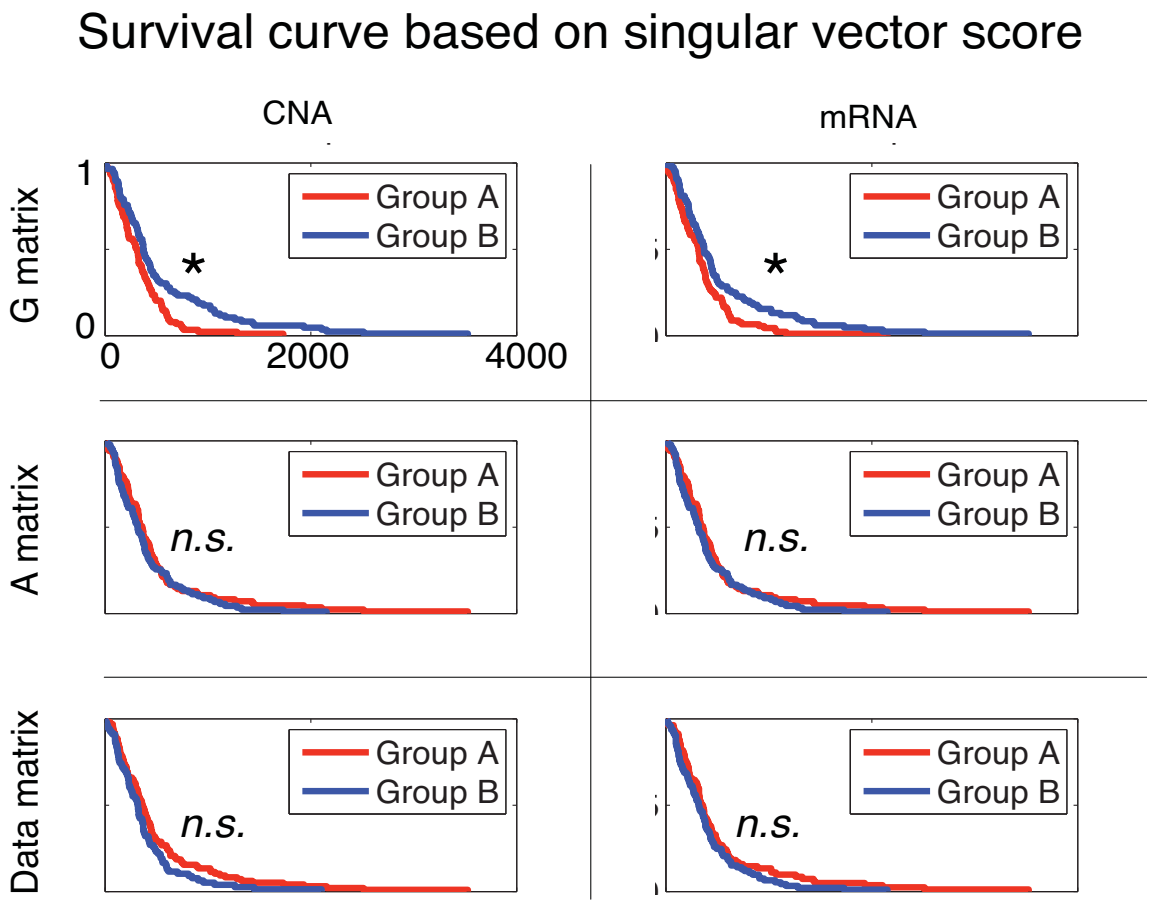
- We argue that G represents the disease mechanism.
- The SVD decomposition of $G = C\Lambda D^T$ has the following meaning:
  - leading columns of D are directions of CNA perturbations that are amplified by the system.
  - leading columns of C are directions of mRNA transcripts most affected by the directions in D.
- Write $Y = GU = C\Lambda D^T U$
- $\rightarrow C^T Y = \Lambda D^T U$
- Projecting mRNA onto columns of C = output, Projecting CNA onto D = input, Λ = amplification.

Decomposition of G



- $C^T Y = \Lambda D^T U$
- Consider the leading projections (first columns of C and D).
- mRNA profiles of individual patients are projected onto C: $Z_y = C^T Y$ and CNA profiles are projected by $Z_u = D^T U$
- Compare the survival of the patients using these projected scores.

Decomposition of G vs A and data



Conclusions

- EPoC scales to 10000 genes and produces stable network estimates
- Attained network models exhibit good agreement with pathway databases
- Experimental validation of novel hubs identify interesting therapeutic targets
- The EPoC network provides clinical stratification into long- and short-term survival, whereas competing methods do not.

Future Work

- Extend EPoC to tumor subtype identification (promising results already)
- Common and subtype specific network modules (ongoing work with PhD students in my group)
- Include multiple data sources (e.g. miRNA, methylation)
- Methodological work - supervised prognostic network estimation

Acknowledgements

- Sven Nelander
- Tobias Abenius
- Teresia Kling, Linnea Schmidt, Bodil Nordlander, Erik Johansson, Torbjörn Nordling, Chris Sander, Björn Nilsson, Peter Gennemark, Keiko Funa, Linda Lindahl
- Cancerfonden, Barncancerfonden, Vetenskapsradet, BioCare, Sahlgrenska-CMR, NB-CNS