

A Principal Component Regression-based Approach for Data-Driven Network Reconstruction

February 11, 2011

Mano Ram Maurya

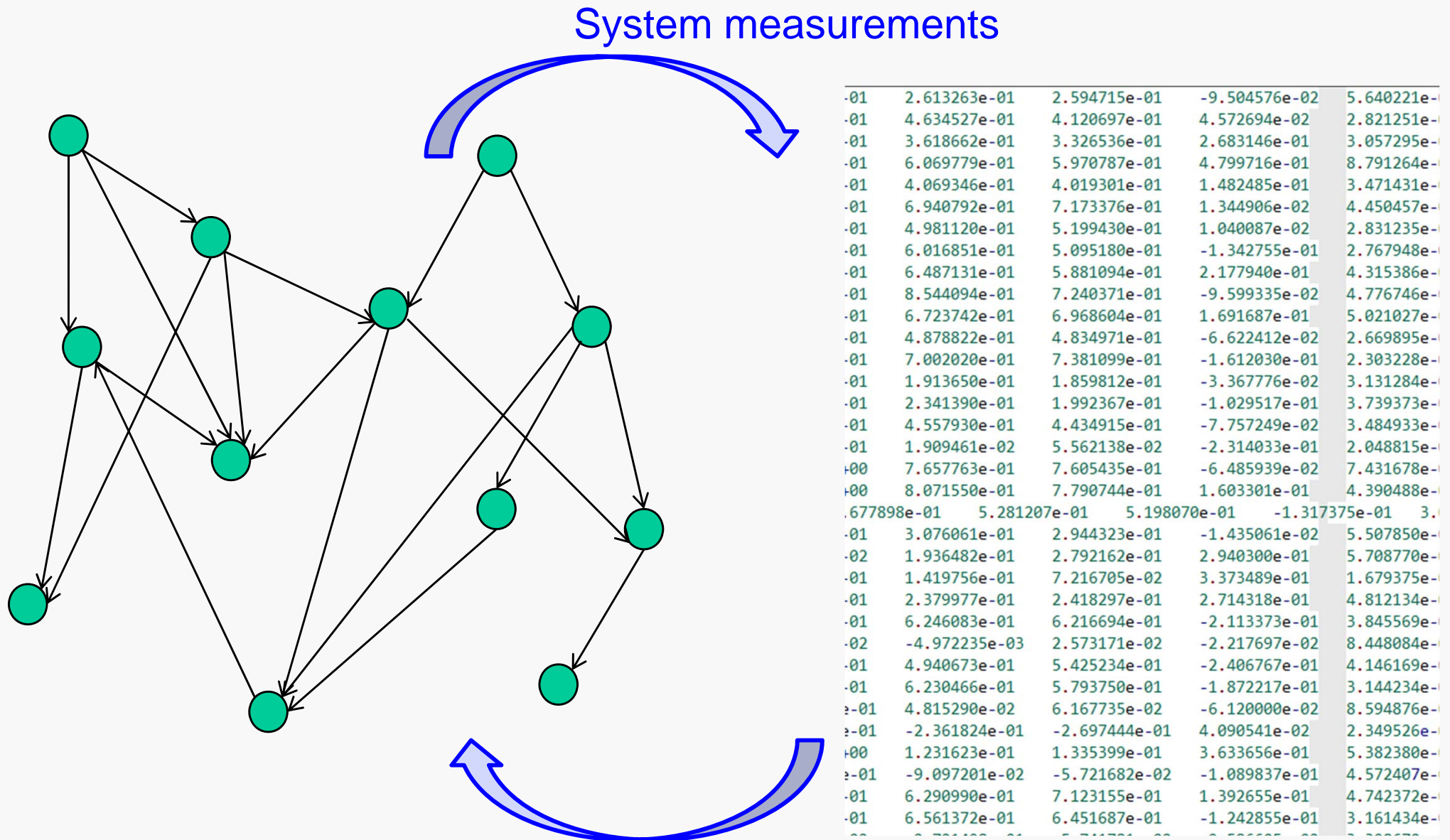
Sylvain Pradervand

Shankar Subramaniam

Outline

- Basics of data-driven network reconstruction
 - ❑ Least-squares approach
- Principal component regression
 - ❑ Basics and statistical significance testing
- Application to phosphoprotein/cytokine modules in macrophages
 - ❑ A two-part model
 - ❑ Results: reduction in false-positive rate
 - ❑ Summary

Network Reconstruction Using Data



How to identify the network back: Data-driven Network Reconstruction

Least Squares Approach

➤ Estimation of b using least squares

$$Y = Xb + E; \quad E \sim N(0, \sigma)$$

$$\hat{b} = (X^T X)^{-1} X^T Y$$

➤ Prediction of output data for the training and test sets

$$\hat{Y} = X\hat{b} = X(X^T X)^{-1} X^T Y$$

$$\hat{Y}_{test} = X_{test}\hat{b}$$

$$E = Y - \hat{Y} = [I - X(X^T X)^{-1} X^T]Y$$

$$E_{test} = Y_{test} - \hat{Y}_{test}$$

$$S_{train} = \sum_i e_i^2$$

$$S_{test} = \sum_j e_{test,j}^2$$

➤ Validation: F-test

- ❑ Separate test set vs. k-fold cross-validation

Outline

- Basics of data-driven network reconstruction
 - Least-squares approach
- Principal component regression
 - Basics and statistical significance testing
- Application to phosphoprotein/cytokine modules in macrophages
 - A two-part model
 - Results: reduction in false-positive rate
 - Summary

Limitation of Least Squares Approach

➤ When X does not have full column rank

- ❑ Columns dependent: $X^T X$ is singular and $(X^T X)^{-1}$ does not exist

$$Y = Xb + E; E \sim N(0, \sigma) \Rightarrow \hat{b} = (X^T X)^{-1} X^T Y$$

➤ When X has less number of rows than columns

- ❑ Less rows than columns: $X^T X$ singular

➤ When the number of rows and columns about the same

- ❑ Least squares may not be reliable

➤ Principal component regression or Partial least squares approach

Principal Component Regression Approach

- Estimate B st. $Y = X*B$, using known X and Y
(generally one output at a time)
- Essence: From X, extract independent components

X: input (predictor data): $m*n$

Y: output: $m*p$ ($p = 1$)

SVD or Eigen-value/vector analysis:

V = matrix of eigen vectors of $\text{cov}(X)$

T = matrix of latent variables

k latent variables (LV) used

X		$Y = X*B$
↓	V	$T = X*V$
T		$Y = T*Q$
↓	Q	$Y = X*V*Q$
Y		$B = V*Q$

1. Calculate Q using least-square method
2. Predict Y: $Y_p = T*Q$

Statistical Significance of the Coefficients

- Most coefficients non-zero
- Identify the significant coefficients
 - ❑ Estimate standard deviation of the coefficients (σ_j)
 - ❑ Approximation (Dash et al. 2004; Pradervand et al., 2006):

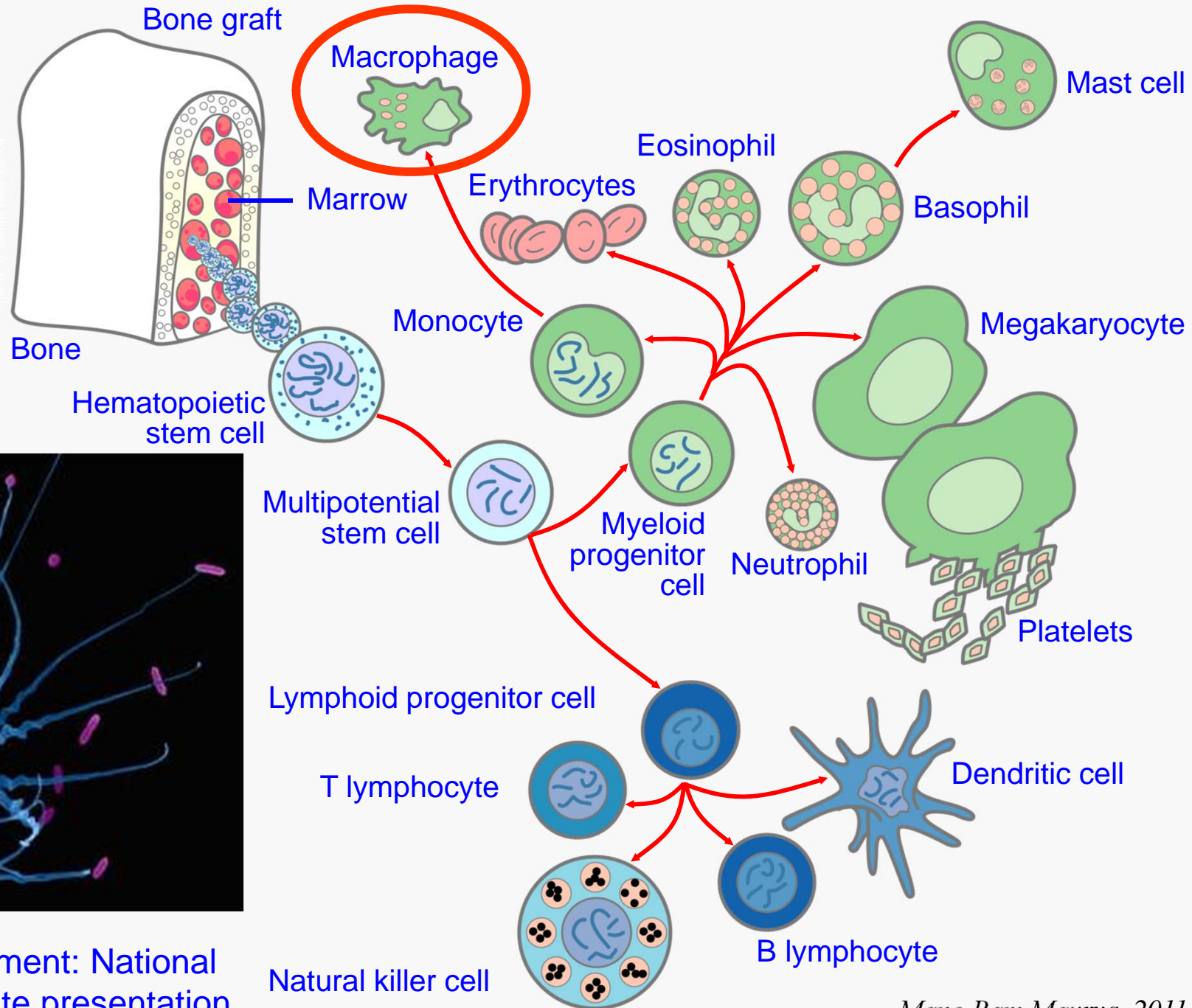
$$\sigma_j \approx \text{diag}(V * (\Lambda_k * (m-1))^{-1} * V^T) * \text{std}(Y_j - Y_{j,p})$$

- ❑ Calculate the ratio: $r_j = b_j / \sigma_j$
- ❑ Significance (t-) test at 95% confidence level: $r_{th} \cong 1.96$
- ❑ Null hypothesis true if average $r_j < r_{th}$

Outline

- Basics of data-driven network reconstruction
 - ❑ Least-squares approach
- Principal component regression
 - ❑ Basics and statistical significance testing
- Application to phosphoprotein/cytokine modules in macrophages [with Pradervand et al., 2006]
 - ❑ A two-part model
 - ❑ Results: reduction in false-positive rate
 - ❑ Summary

Macrophages: Cells of the Immune System



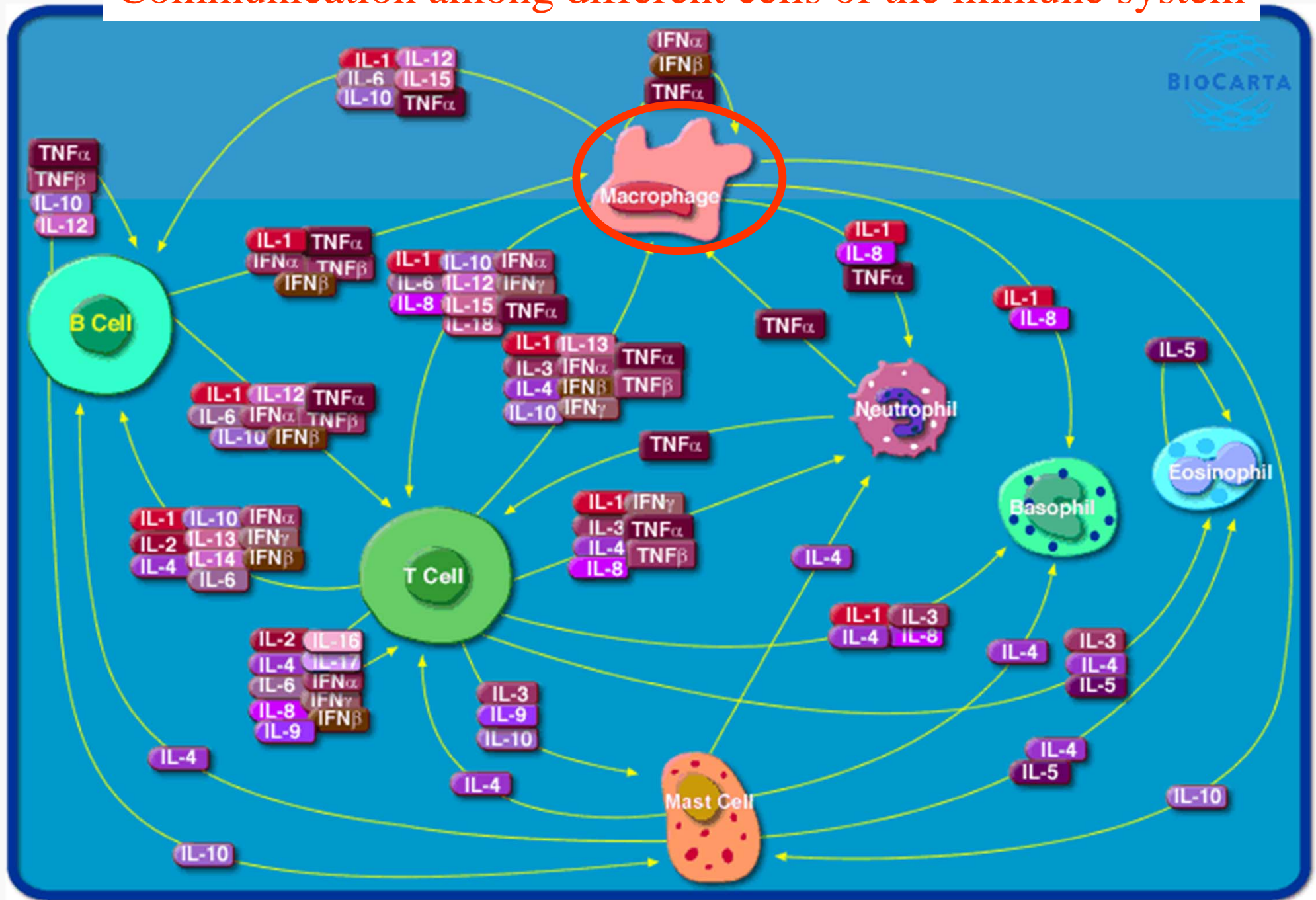
Acknowledgement: National Cancer Institute presentation

Cytokines: The Communicator Molecules

➤ Cytokines

- ❑ Proteins made in T and B cells and macrophages
- ❑ facilitate communication between immune cells

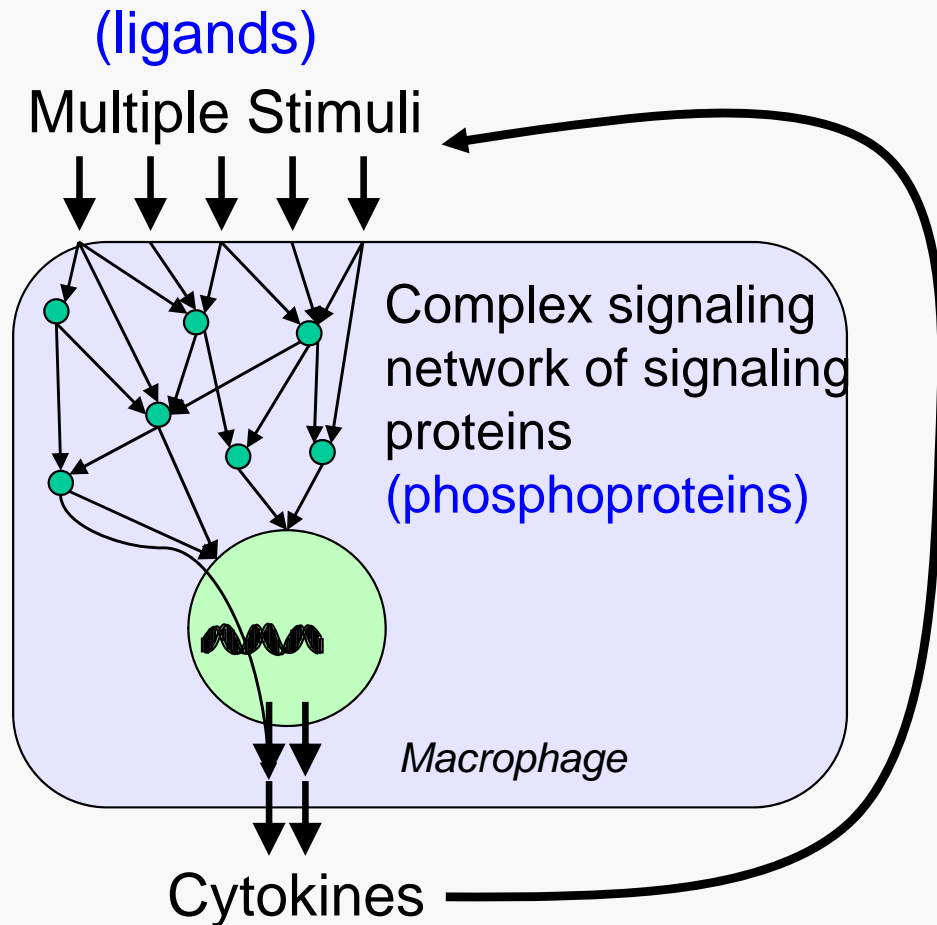
Communication among different cells of the immune system



Objective

- Each cytokine regulated by several distinct signaling pathways
- Many known but a lot is unknown
- **Questions**
 - ❑ **Can we find out which signaling pathways control a specific cytokine using high-throughput data and computational approaches?**

Cytokine Production and Release in Macrophages

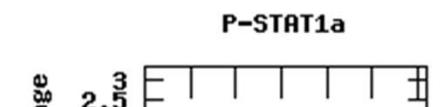
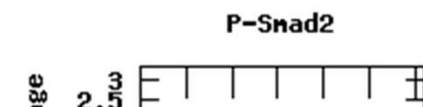
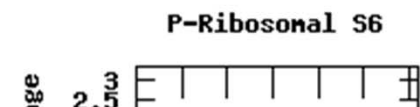
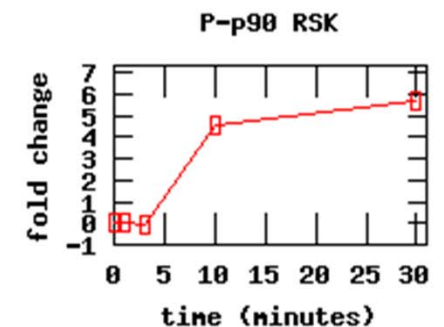
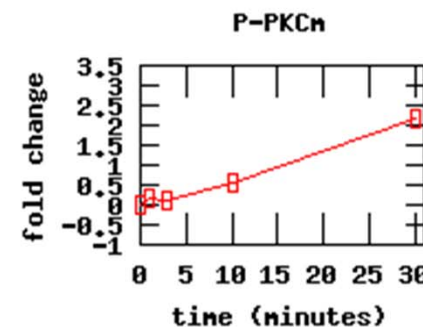
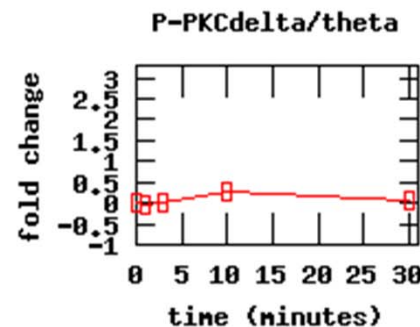
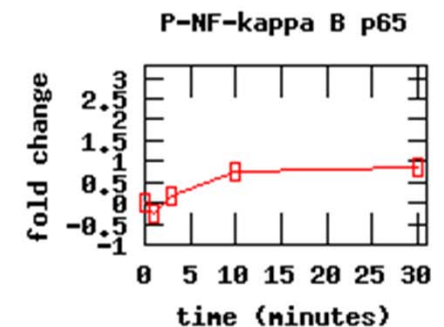
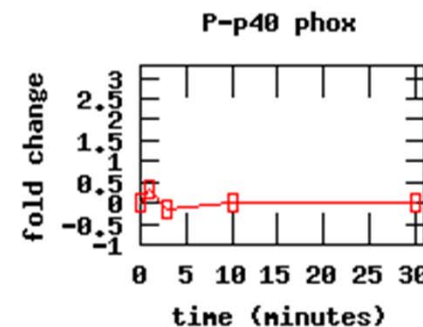
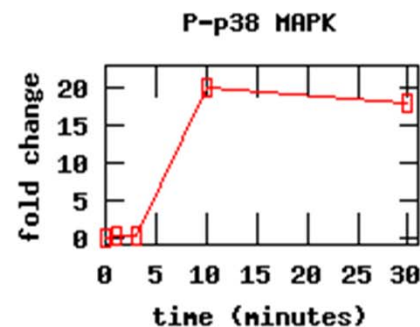
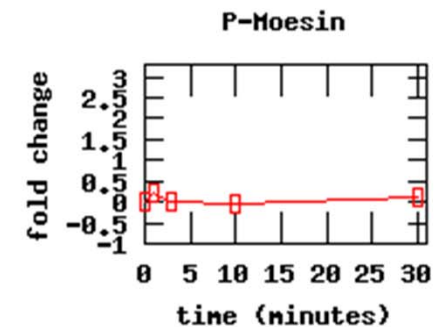
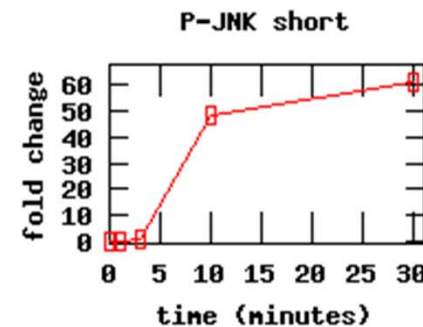
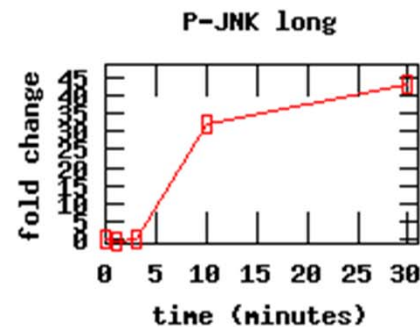


- Data from the Alliance for Cellular Signaling (AfCS)
- ~ 250 experiments
 - ❑ 22 signaling pathway markers (inputs)
 - ❑ 7 cytokines (outputs)

Experimental Data: Phosphoproteins

Measurements of phosphoproteins in response to LPS

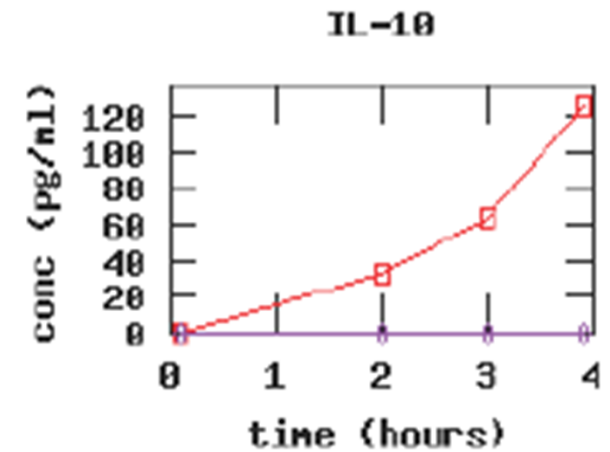
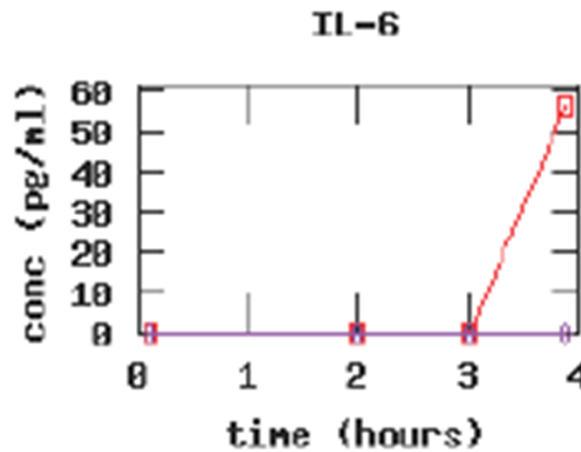
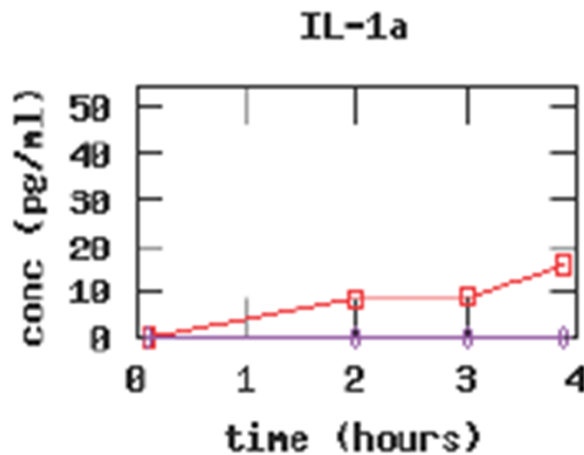
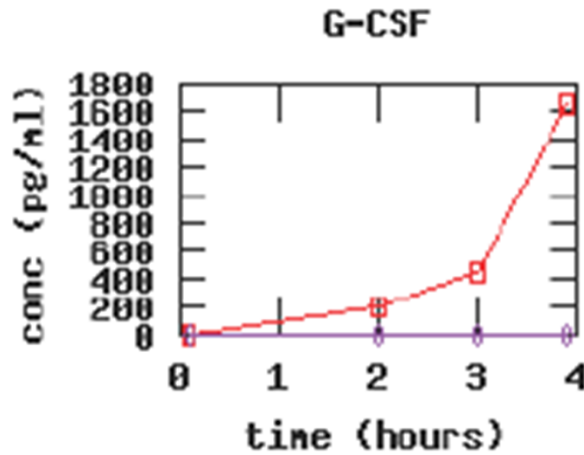
	UDP	TGF	S1P	PGE	PAF
2-Methyl-thio-ATP	Y	Y	Y	N	Y
Resiquimod (R-848)	Y	Y	Y	Y	Y
Complement C5a	N	N	Y	N	Y
Granulocyte macrophage colony stimulating factor	Y	Y	N	Y	N
Interleukin-4	Y	Y	N	Y	Y
Interleukin-6	Y	N	Y	N	Y
Interleukin-10	Y	Y	N	Y	Y
Interleukin-1-beta	Y	Y	N	Y	Y
Interferon-alpha	Y	N	Y	Y	Y
Interferon-beta	Y	Y	Y	Y	Y
Interferon-gamma	Y	Y	Y	N	Y
Isoproterenol	Y	N	Y	Y	Y
Lysophosphatidic acid	N	i	Y	Y	N
Lipopolysaccharide	Y	Y	Y	N	Y
Macrophage colony stimulating factor	Y	Y	Y	N	Y
PAM2CSK4	Y	Y	Y	N	Y
PAM3CSK4	Y	N	Y	Y	Y
Platelet Activating Factor	Y	Y	Y	Y	
Prostaglandin E2	Y	N	Y		
Sphingosine-1-phosphate	Y	N			
Transforming growth factor-beta	Y				



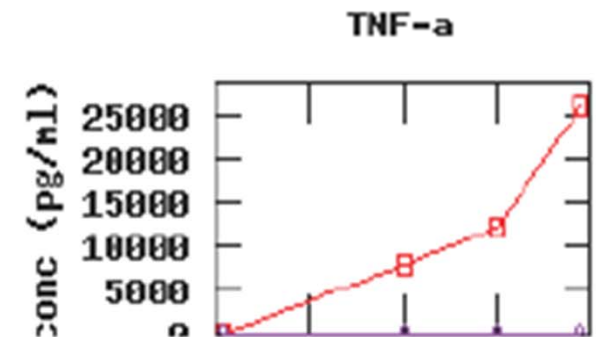
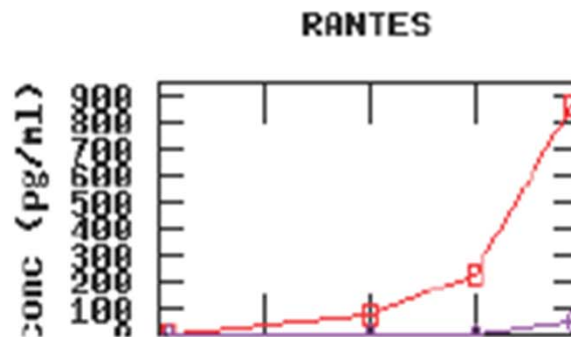
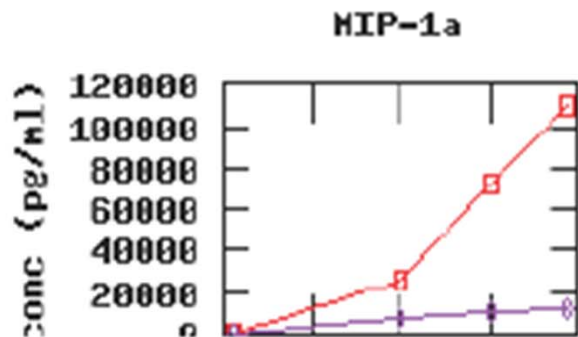
Courtesy: AfCS

Experimental Data: Cytokines

Measurements of cytokines in response to LPS



~ 250 such datasets



Courtesy: AfCS

Structure of the Model

➤ Two-part model

➤ Part-I: Capture most of the output as:

$$Y_1 = X * \mathbf{B}_1$$

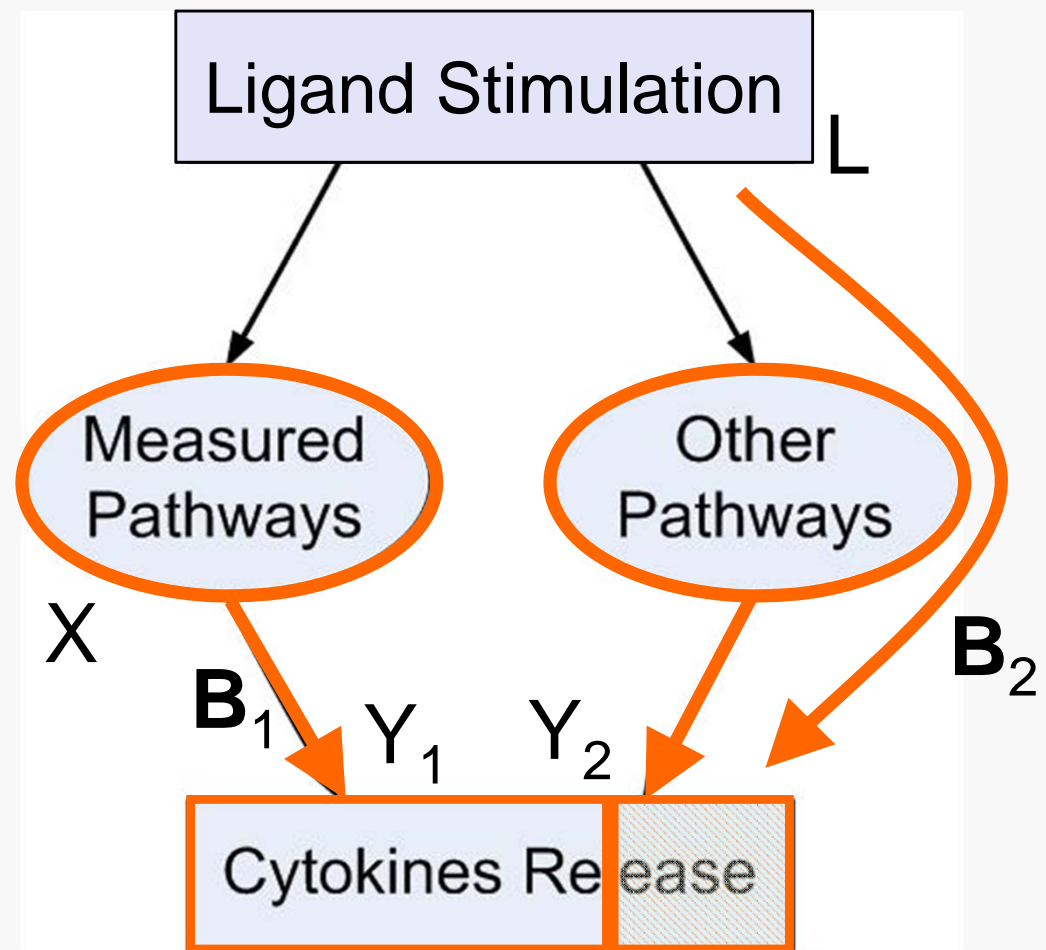
- Threshold on the ratio,

$$r_j = b_j / \sigma_j : r_{th} = 1.96$$

➤ Part-II: Residual

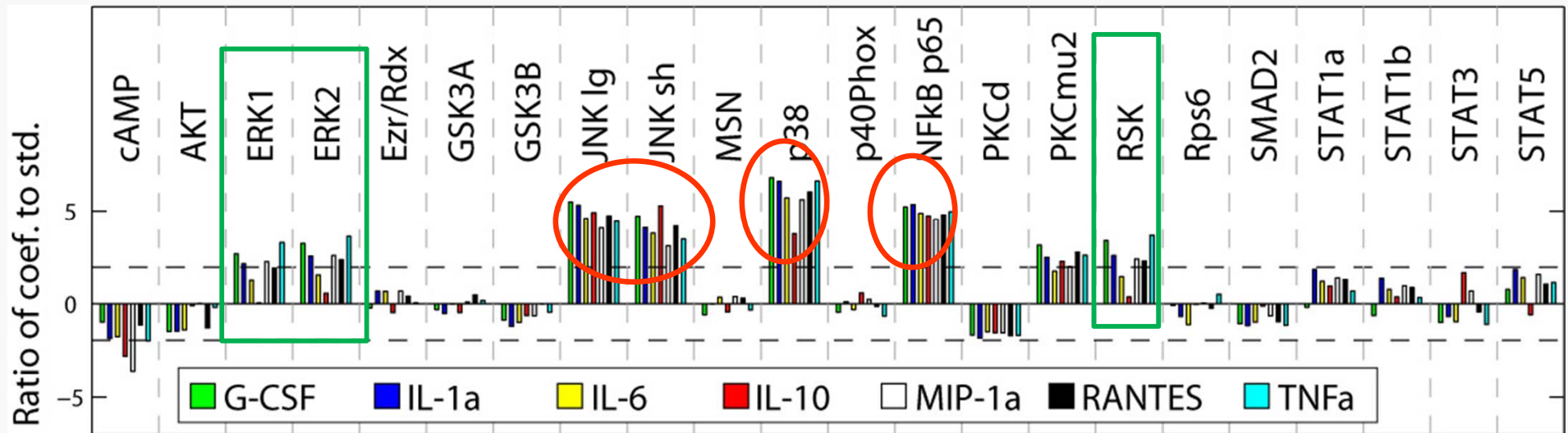
$$Y - Y_1 = Y_2 = L * \mathbf{B}_2$$

- $r_{th} = \sqrt{2} * 1.96 = 2.77$



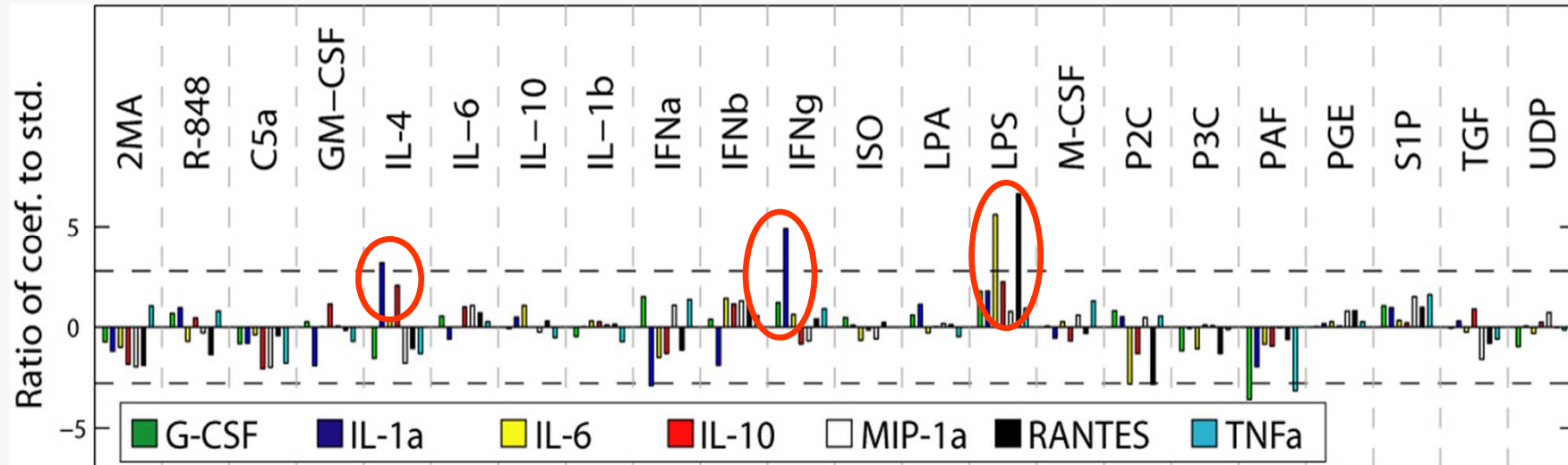
$$\text{Total cytokine: } Y = Y_1 + Y_2$$

Results: Insights into Pathways



- JNK, p38, NF-kB strongest coefficients
 - ❑ Exert substantial control on transcription
 - ❑ A candidate for drug targets for inflammatory diseases
- ERK1/2 and RSK (ribosomal S6 kinase) similar profile
- cAMP is anti-correlated

Significant Unmeasured Pathways



- Just few ligands significant
- IL-4 and IFN-g for IL-1a
- LPS for IL-6 and RANTES: possible role of non-canonical (other than NF-kB) pathways

Minimal PCR Model

- Due to correlations in the inputs (X), high false positive rate
- Identify a minimal subset of the signaling pathways

Procedure for Minimal PCR Models

F-test:

As good as the full-model:

$$R_1 = e_r^2 / e_d^2 < finv(p, d_r, d_d)$$

$$p = 1 - \alpha, \alpha = 0.05, p = 0.95$$

Better than the trivial model:

$$R_2 = e_0^2 / e_r^2 > finv(p, d_0, d_r)$$

If more than one predictor left, use combinatorial selection for exhaustive testing

Decreasing number of predictors

Full (detailed) model with all significant predictors (e_d)

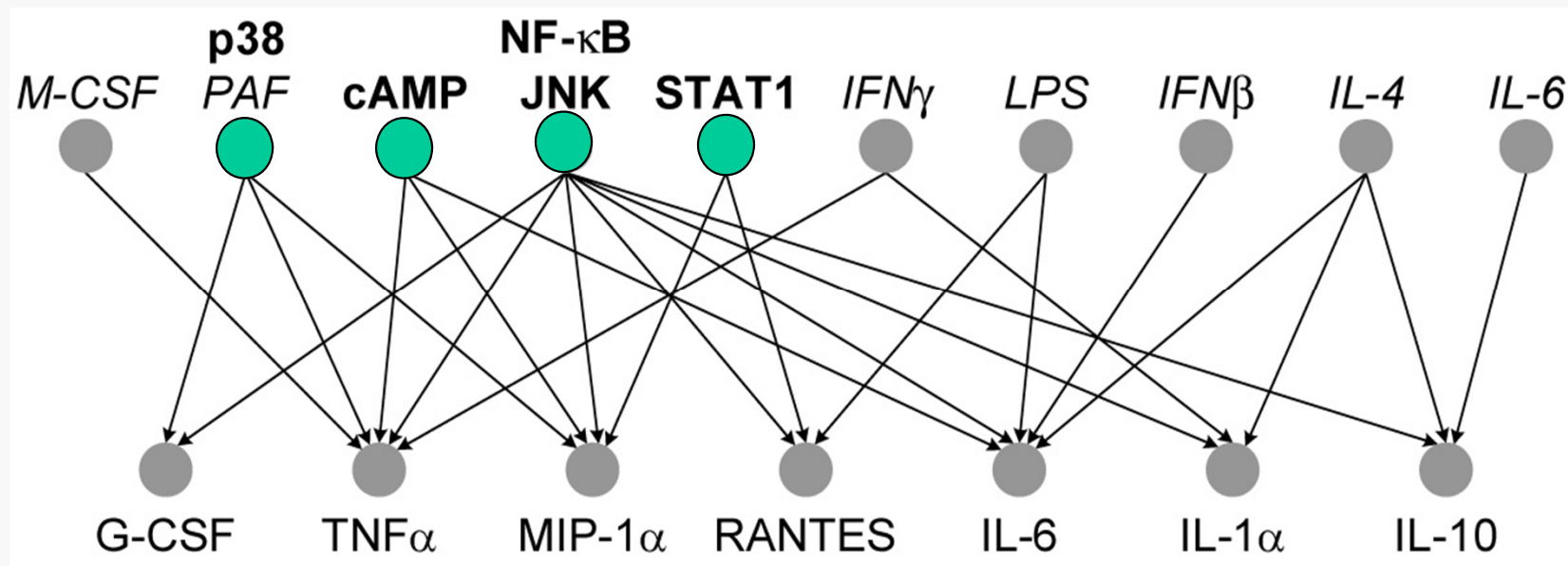
Keep eliminating the least significant predictor:
 R_1 increases, R_2 decreases

Initial minimal model

Final minimal model

Zero-predictor model (e_0)

Combined Minimal Model Reveals Modules



A combined global network map

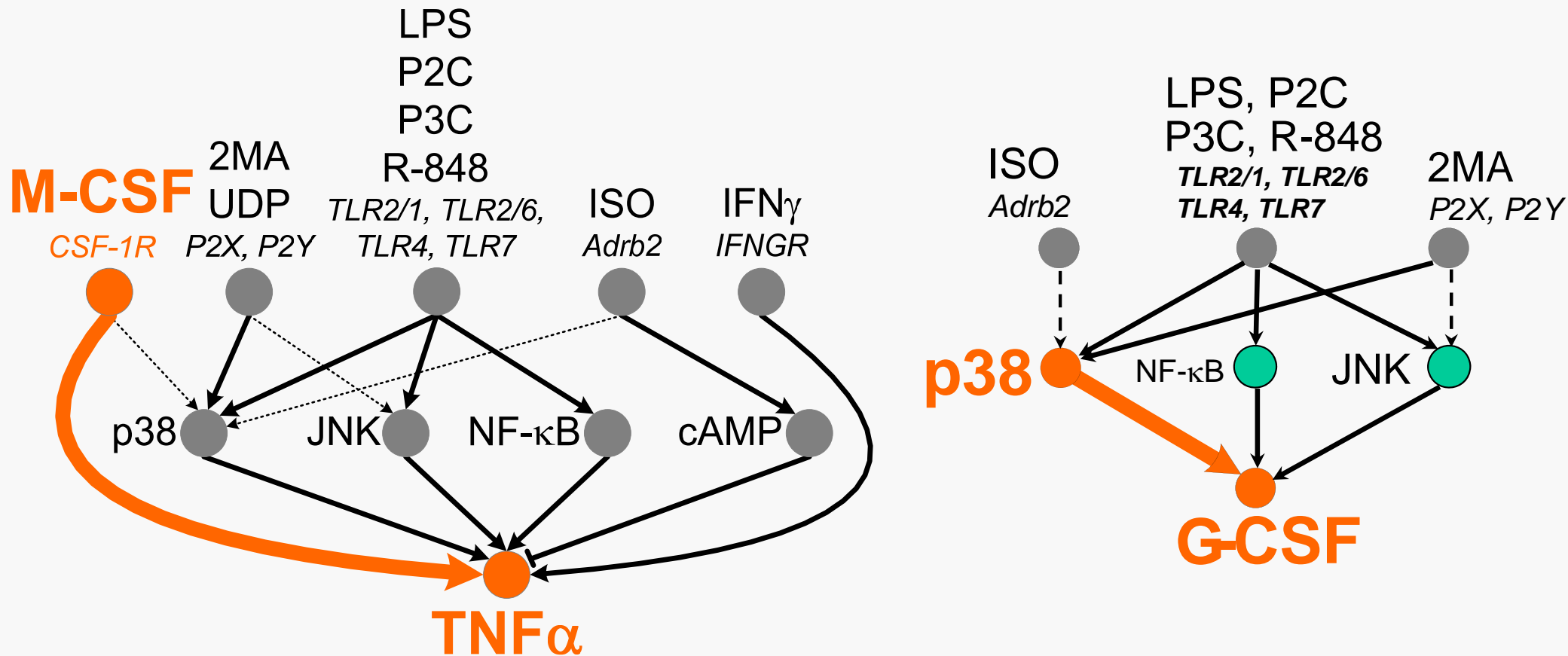
□ 10 regulatory modules

Measured Pathways: p38, cAMP, NF-kB, JNK, STAT1 (4)

Ligands: others (6)

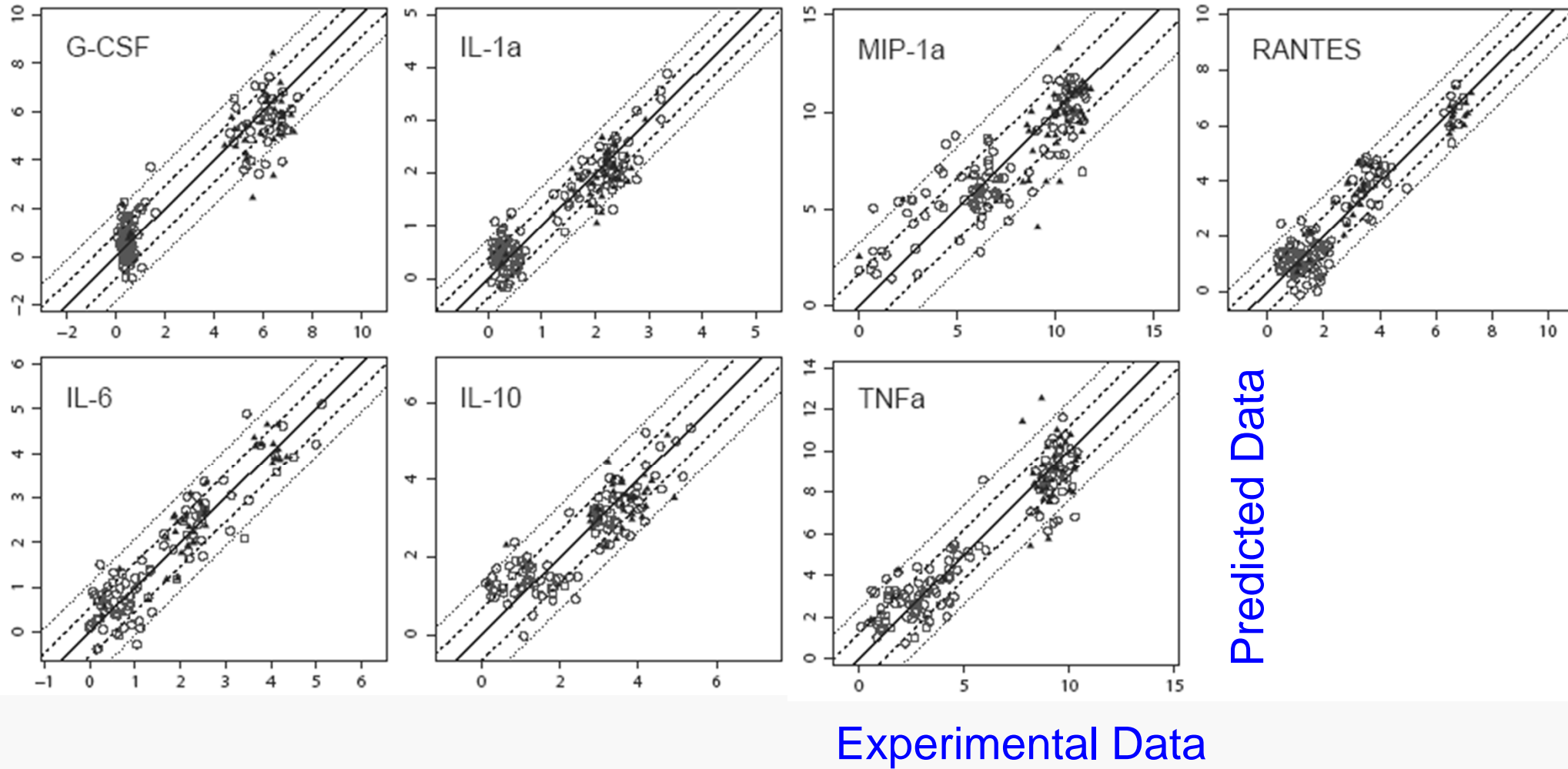
Pradervand, Maurya et al., *Genome Biology*, 2006

Results: Novel Hypothesis



- $TNF\alpha$: a target for Rheumatoid Arthritis
- New hypothesis: M-CSF-specific pathway regulates $TNF\alpha$?

Validation: Prediction of Test Data



Validation with the Literature

- Found most known pathways
- Substantial decrease in the number of false positives
- Critically important in target discovery due to high cost of target screening

	False Positive Rate	False Negative Rate
Full model	11% (28 false connections)	3% (1 missing)
Minimal model	1.2% (only 3 false connections)	13% (5 missing)

Summary

- Coarse-grained network maps for cytokine release
- Data-driven network reconstruction
 - ❑ Provides quantitative insights into systemic regulation
 - ❑ Discover novel regulatory pathways
- Aid in designing new approaches to drug discovery

Acknowledgements

- Various grants from the NSF and NIH
- AfCS for experimental data



