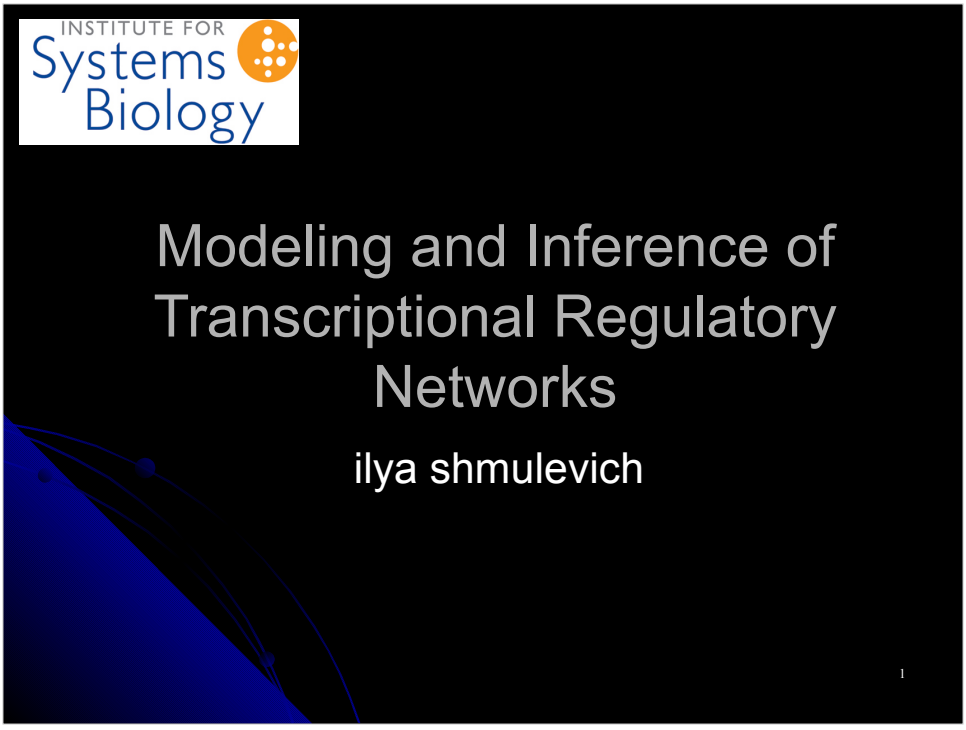
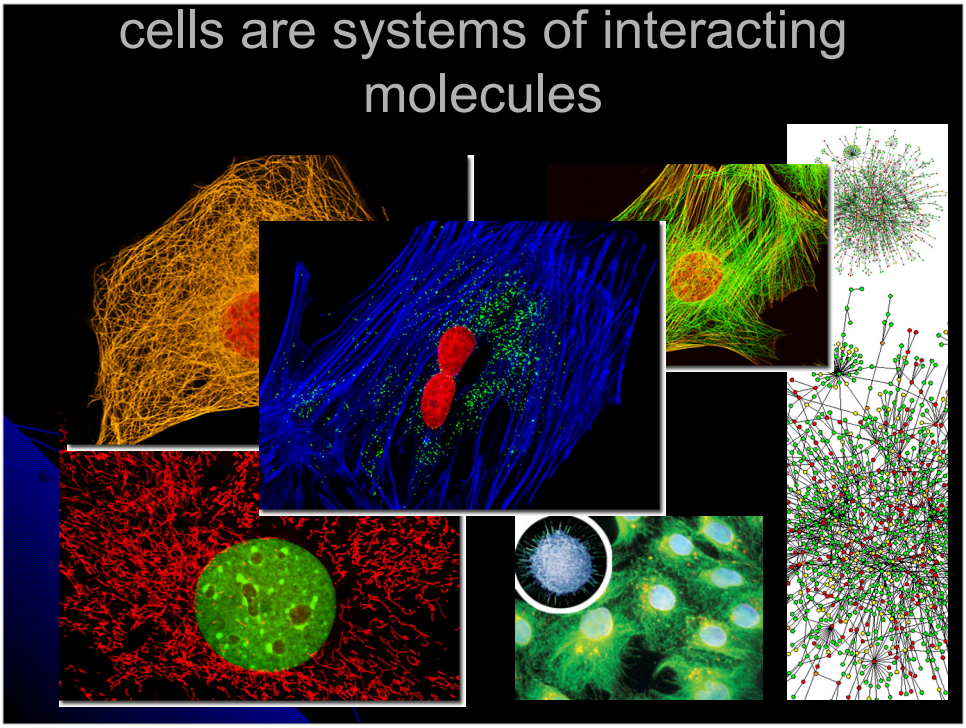
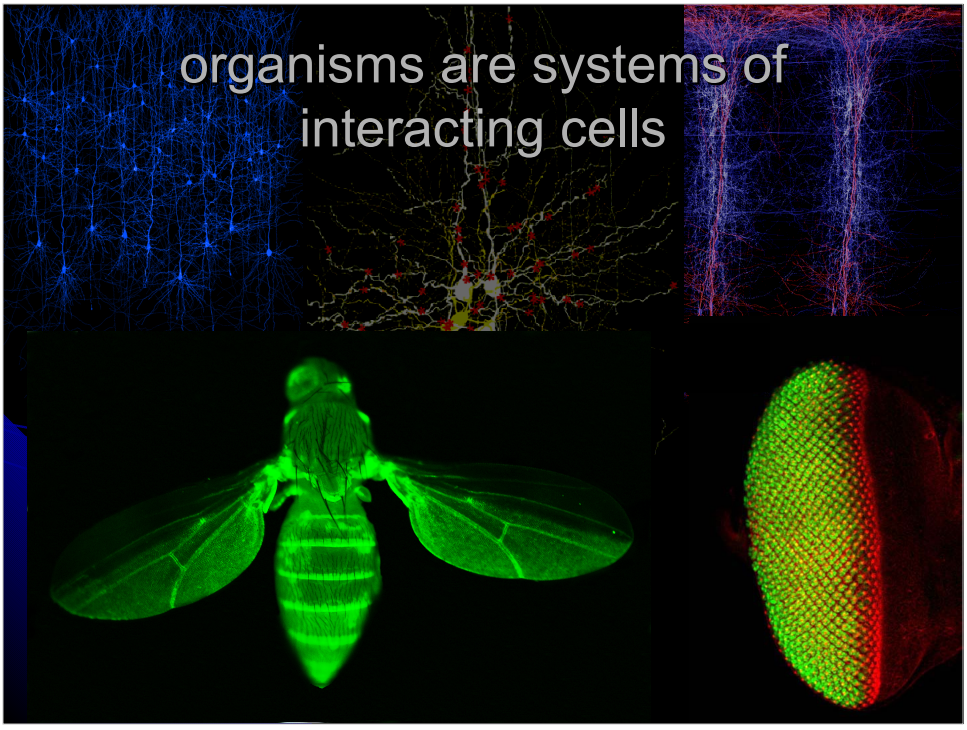

INSTITUTE FOR
Systems
Biology
Modeling and Inference of Transcriptional Regulatory Networks
ilya shmulevich
1


cells are systems of interacting molecules

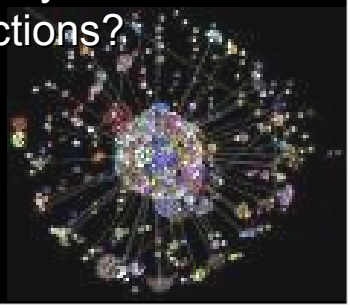

organisms are systems of
interacting cells

societies are systems of interacting organisms

living systems

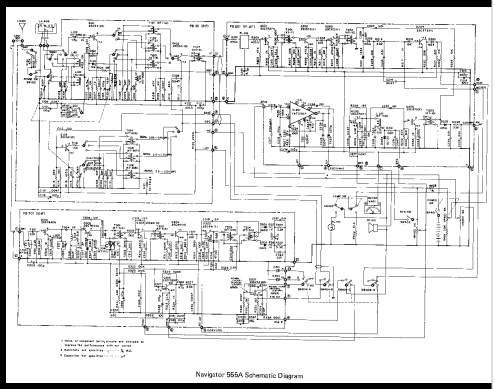
self-organized complex dynamical systems of interacting parts

how can we understand the emergent macroscopic properties of the system from its parts and their interactions?



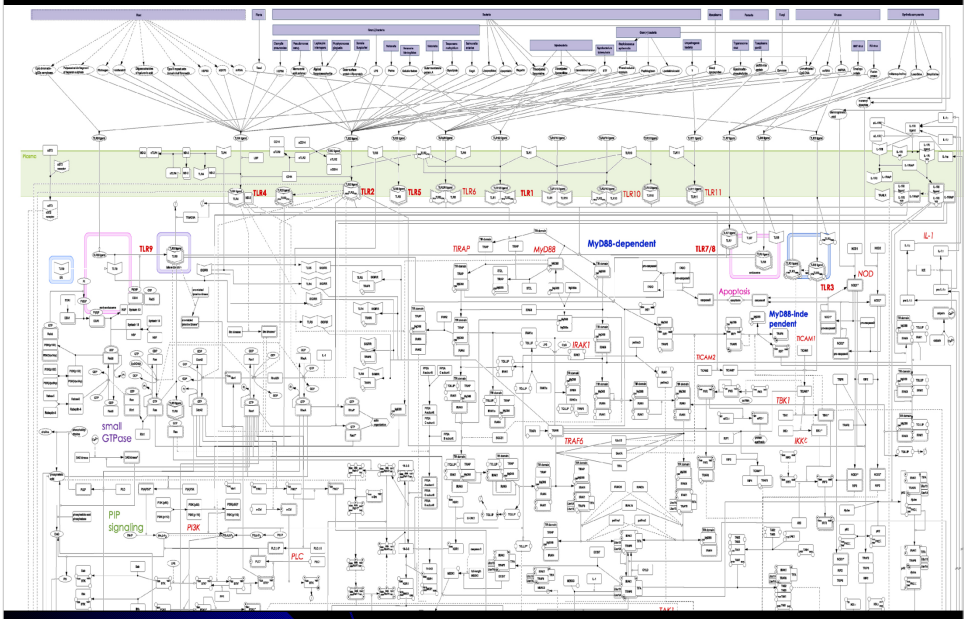
genetic networks

- Complex regulatory networks among genes and their products control cell behaviors such as:
  - cell cycle
  - apoptosis
  - cell differentiation
  - communication between cells in tissues
- A paramount problem is to **understand the dynamical interactions** among these genes, transcription factors, and signaling cascades, which govern the integrated behavior of the cell.



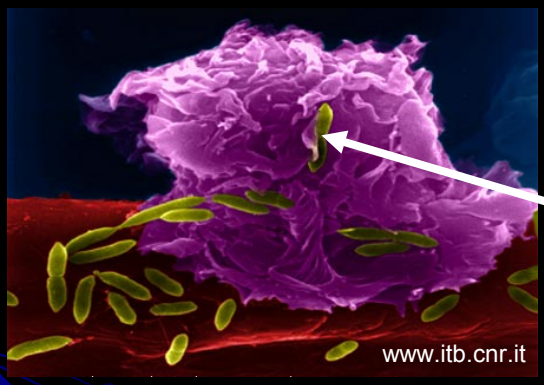
Analogy: circuit diagram

Map of the TLR signaling pathway in the macrophage



Mol Syst Biol. 2:2006.0015

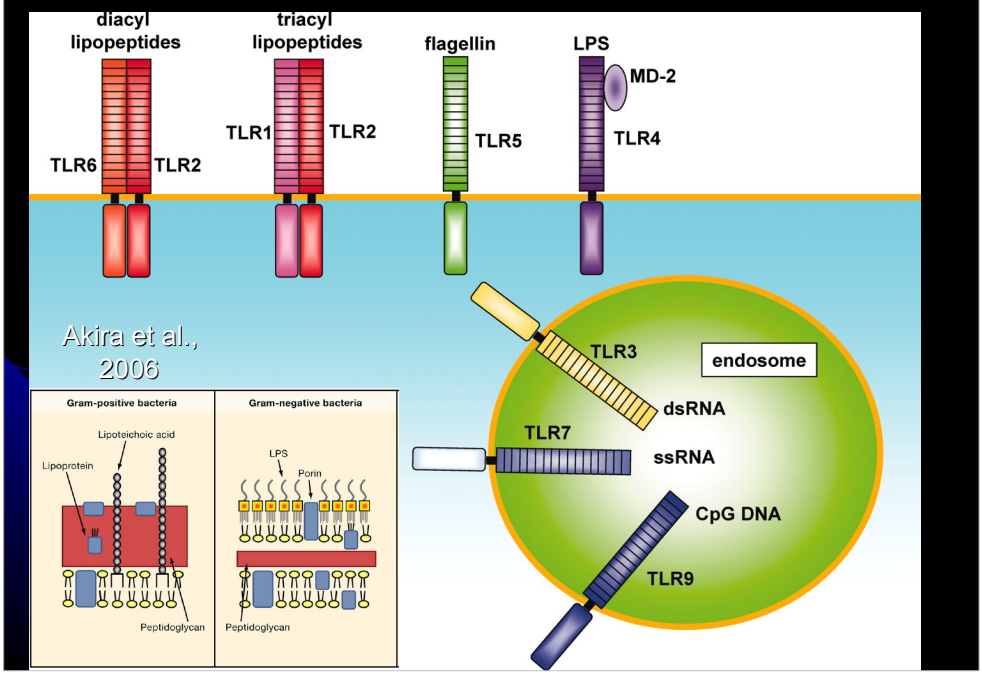
Macrophages are key immune cells



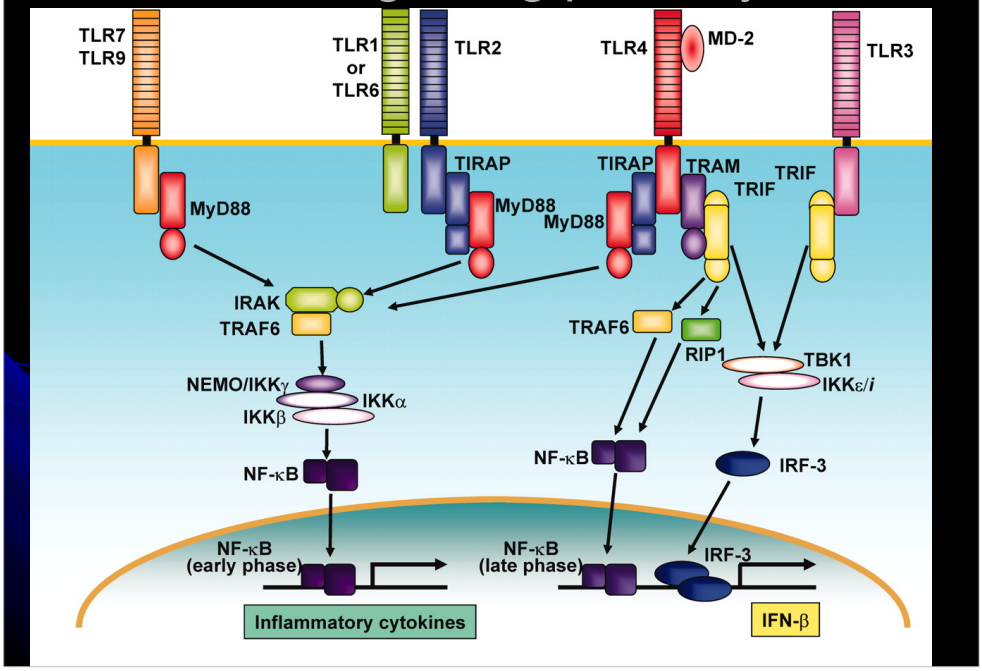
bacterium
being engulfed

- Phagocytosis
- Antigen presentation
- Secretion of proinflammatory cytokines
- Wound healing

TLRs and their associated PAMPs

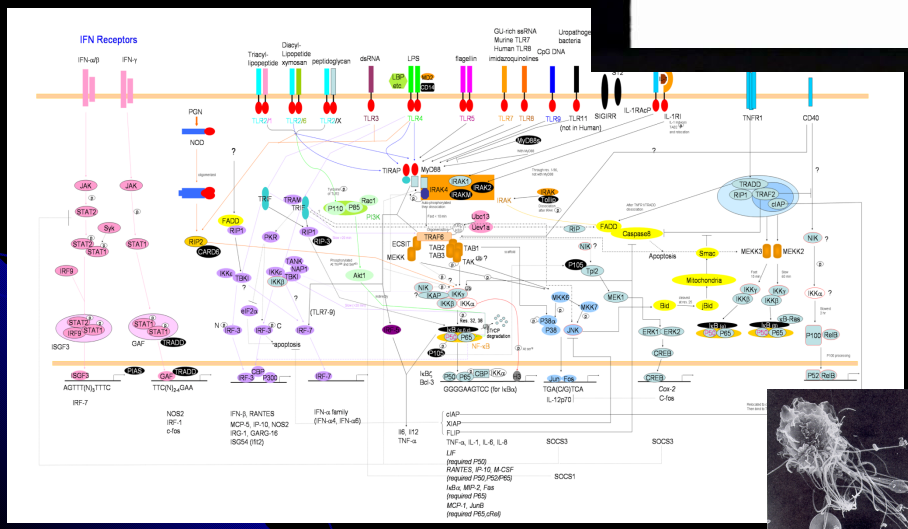


TLR signaling pathway



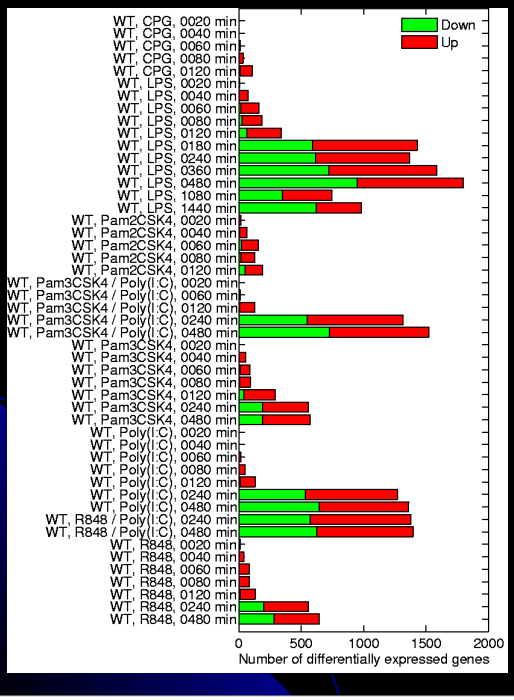
Map of the TLR signaling pathway in the macrophage

A few perspectives on
CELL MOTILITY
Division of Experimental Medicine
Brigham & Women's Hospital



The ingredients

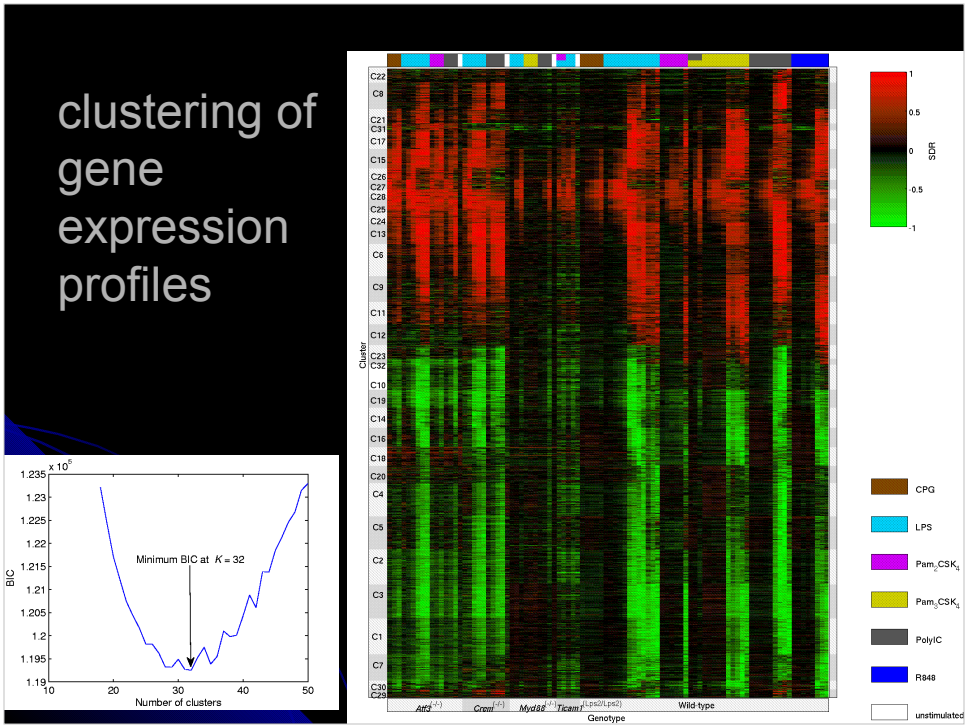
- 130 microarray experiments (253 arrays)
  - Seven mouse strains (WT, Ahr$^{(-/-)}$, Atf3$^{(-/-)}$, Crem$^{(-/-)}$, Cebpd$^{(-/-)}$, Myd88$^{(-/-)}$, Ticam1$^{(Lps2/Lps2)}$)
  - Combinations of six stimuli (LPS, Pam_3CSK_4, Pam_2CSK_4, poly I:C, CpG, R848, T091317)
  - Time courses out to 8 hours (24 for LPS)
- Mouse genome promoters (UCSC)
- TRANSFAC Professional 10.3
- Curated list of ~1800 human TFs



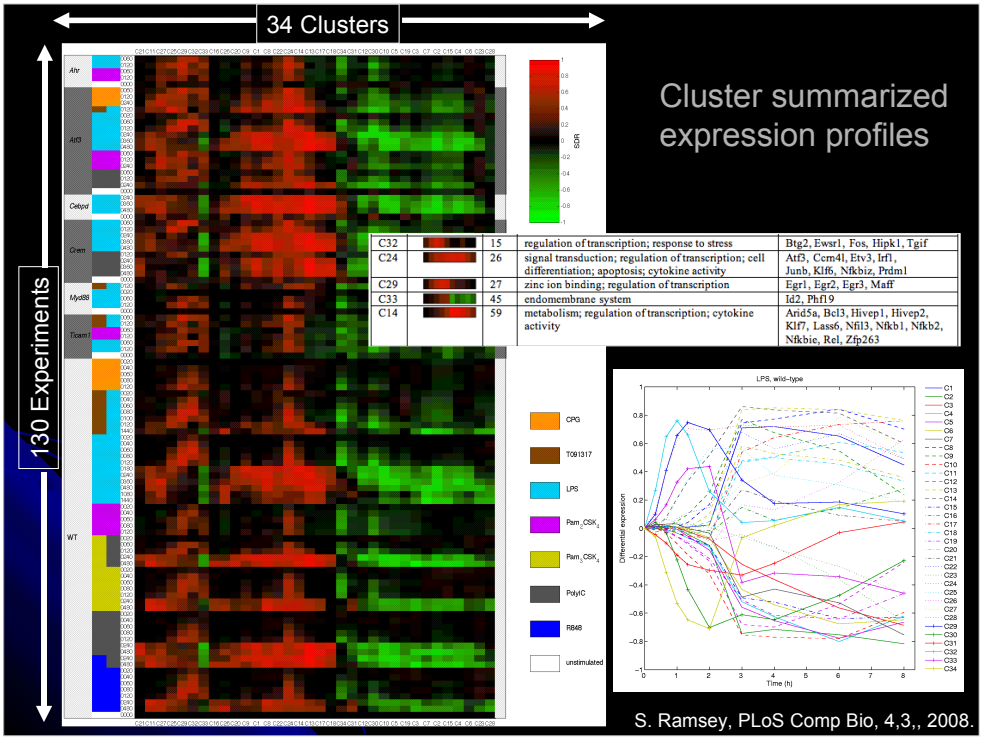
Number of differentially expressed genes vs. elapsed time (by stimulus)

Total number: 2,562

S. Ramsey, PLoS Comp Bio, 4,3, 2008.

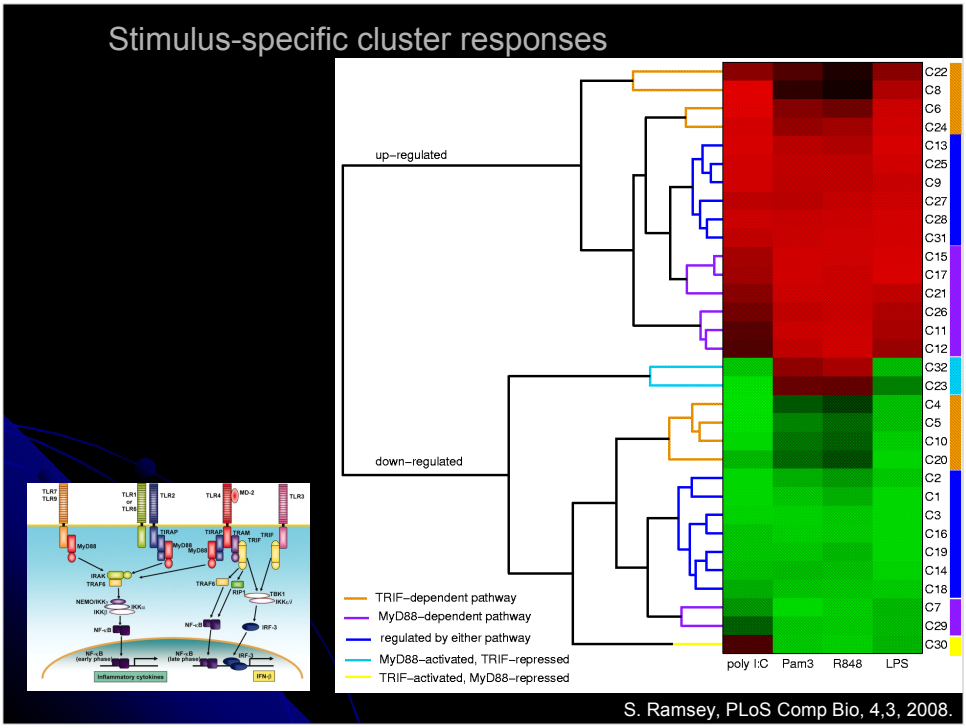


We used K-means clustering of normalized differential expression levels of genes, to identify clusters of genes with similar temporal and stimulus-specific patterns of differential expression (relative to wild-type unstimulated macrophages). The number of clusters was varied, and we selected 32 clusters because it best described the data (minimized the Bayesian Information Criterion). Red indicates upregulated genes, green indicates downregulated genes. Each row is a gene, and each column is a microarray experiment (216 hybs total, corresponding to 95 different sample conditions). Data were from BL/6 macrophages including wild-type and homozygous-null mutants for ATF3, CREM, MyD88, and TRIF. The colors at the top of the heatmap indicate the stimulus (or combination of stimuli). Clusters are ordered so that similar clusters are adjacent (i.e., they are ordered for display so as to minimize the total distance between adjacent clusters).



C32		15	regulation of transcription; response to stress	Btg2, Ewsr1, Fos, Hipk1, Tgif
C24		26	signal transduction; regulation of transcription; cell differentiation; apoptosis; cytokine activity	Atf3, Ccrn4l, Etv3, Irf1, Junb, Klf6, Nfkbiz, Prdm1
C29		27	zinc ion binding; regulation of transcription	Egr1, Egr2, Egr3, Maff
C33		45	endomembrane system	Id2, Phf19
C14		59	metabolism; regulation of transcription; cytokine activity	Arid5a, Bcl3, Hivep1, Hivep2, Klf7, Lass6, Nfil3, Nfkb1, Nfkb2, Nfkbie, Rel, Zfp263

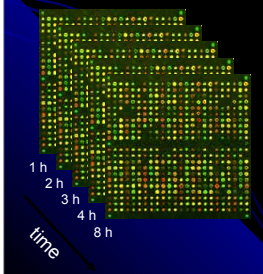
S. Ramsey, PLoS Comp Bio, 4,3,, 2008.



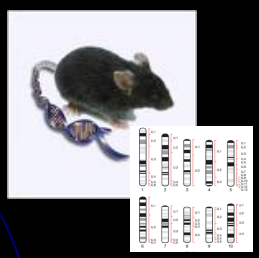
Clusters can also be organized hierarchically based on the extremal differential expression achieved (over time) for each cluster within each of the four stimui for which we have 8-hour time-course data available (poly I:C, Pam3CSK4, R848, and LPS). By comparing the response to the different stimuli, we can formulate a hypothesis about the signaling pathway (or pathways) through which the different clusters may be transcriptionally regulated. These hypotheses are summarized in the colored bars on the right-hand side.

In conjunction with transcription factor binding site prediction, we can use the *timing* of expression to identify induced transcription factors that are associated with downstream groups of genes that they regulate.

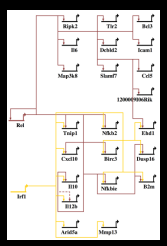
Expression *dynamics*



Genomic sequence data



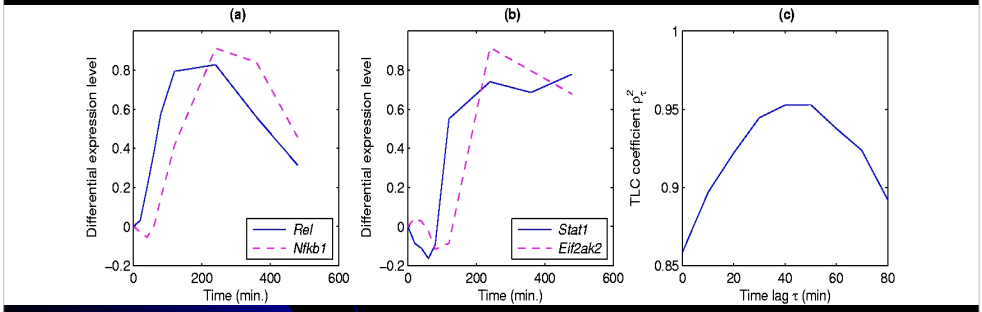
Transcriptional network



S. Ramsey

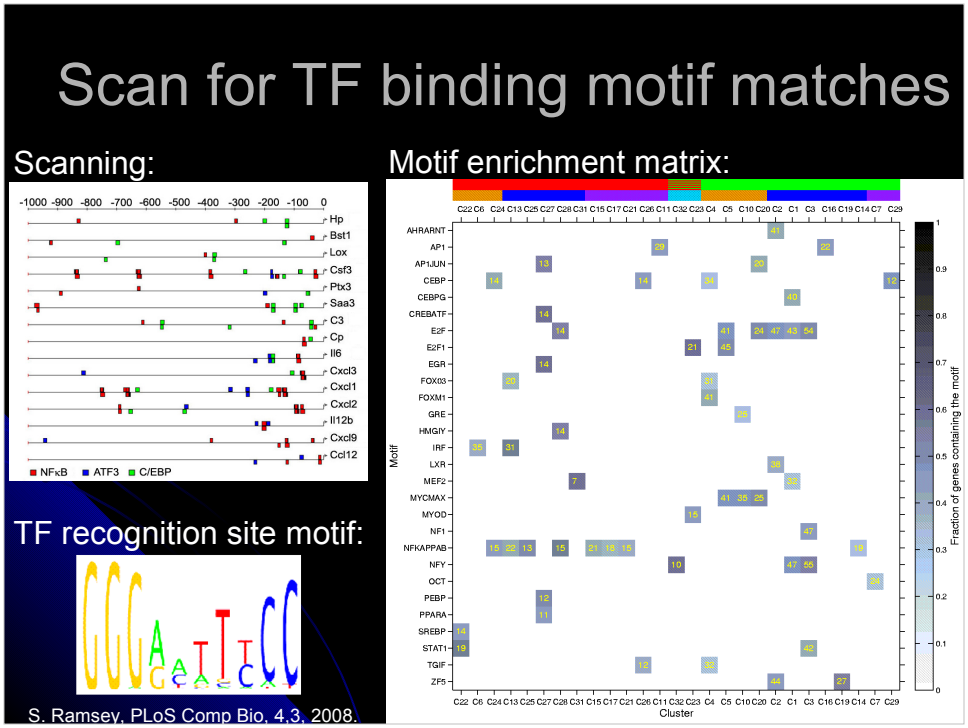
Components of the gene-gene transcriptional time delay

- mRNA half-life of TF gene
- Translation and folding of TF protein
- Diffusion of TF back into nucleus
- Turnover rate (half-life) for TF protein
- Transcriptional elongation of target gene



S. Ramsey, PLoS Comp Bio, 4,3, 2008.

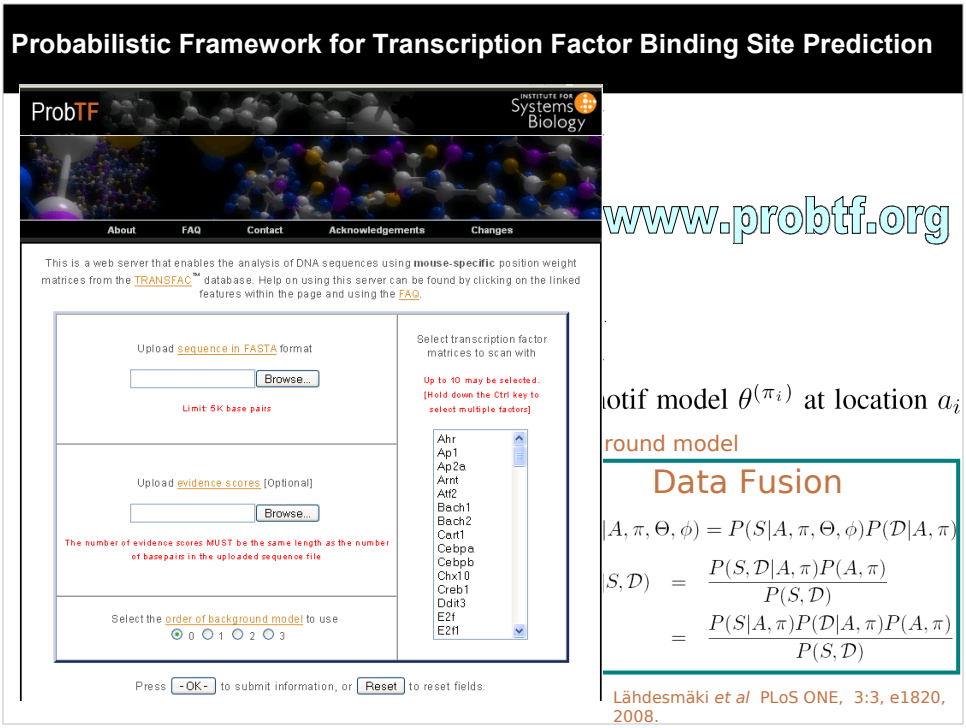
Two validated transcriptional regulatory interactions exhibiting high time-lagged correlations. (a) Rel and Nfkb1. The solid line shows the expression of Rel (c-REL), and the dotted line shows the expression of Nfkb1 (p50 / p105) in LPS-stimulated wild-type macrophages, over eight hours. The genes exhibit a high time-lagged correlation with a time delay of 50 minutes (across the ten time-course experiments listed in Table 6, ρτ = 0.953 and P = 2 × 10-5; see Materials and Methods for an explanation of the statistical test). The NFκB heterodimers c-REL J p50 and c-REL □ p65 are known to regulate expression of Nfkb1 [39]. The correlation at zero time shift is 0.859. (b) Stat1 and Eif2ak2. The solid line shows the expression of Stat1 (STAT1) and the dotted line shows the expression of Eif2ak2 (PKR) in LPS-stimulated wild-type macrophages. The genes exhibit a high time-lagged correlation with a time delay of 50 minutes (across the ten experiments, ρτ = 0.924 and P = 1.2 × 10-3). The heteromeric transcription factor ISGF3 (involving STAT1, STAT2, and IRF9) is known to bind the promoter of Eif2ak2, and that type 1 interferon induction of Eif2ak2 is STAT1-dependent [40]. The correlation at zero time shift was 0.876. (c) Time-lagged correlation coefficient 2 as a function of the time lag τ, for Rel and Nfkb1. The peak occurs at 50 min, and in this case, the optimal time lag * is also at 50 min.



We take clusters of co-expressed genes and scan them to identify matches to pre-computed libraries of motifs (lower left) representing recognition sites of various transcription factors. Upper left: scanning a cluster of PAM3-responsive genes for motifs representing recognition sites for NFkB, CREB/ATF, and C/EBP.

Upper right:

Patterns of high-confidence motif enrichments within promoters of target clusters reveal associations between regulatory elements and expression patterns. Each row in the matrix represents a TF binding element, and each column represents a cluster of differentially expressed genes. Clusters are ordered as in Figure 2, and thus are grouped hierarchically by similarity of their extremal expression fold- change under the four TLR agonists LPS, Pam3CSK4, poly I:C, and R848. Each motif (row) is associated with one or more position-weight matrices (the V$ prefix and numeric suffixes are omitted, and results for multiple position-weight matrices representing the same motif were combined for each column, by taking the matrix with the maximum number of matches within the indicated cluster). Each colored block in the matrix indicates pair of a motif and target cluster for which the fraction of genes in the cluster with a motif match, is enriched relative to the overall fraction of genes expressed in the macrophage that possess the motif (P#1022, Fisher s exact test). The color of each matrix element (block) in the interior of the figure indicates the fraction scanned of genes within the cluster containing at least one match for the indicated motif. The number of scanned genes within the cluster that contained a match for the indicated motif is shown in yellow typeface. The red/green colored blocks above the top horizontal axis shows whether each cluster is upregulated (red) or downregulated (green) at its most extremal fold-change under stimulation with the aforementioned TLR agonists. The hatched green/red pattern indicates a cluster whose extremal fold-change direction (up/down) is stimulus- dependent (see Figure 2). The colored (blue, cyan, orange, yellow, purple) blocks above the top of the matrix indicate the likely pathway through which the cluster is differentially expressed; the color scheme corresponds to that shown in the dendrogram in Figure 2.



Motifs are modeled using the standard PWM model \theta which is a product of independent multinomial distributions.

A natural way to improve specificity of TFBS predictions is to make use of additional information in the inference.

We use:

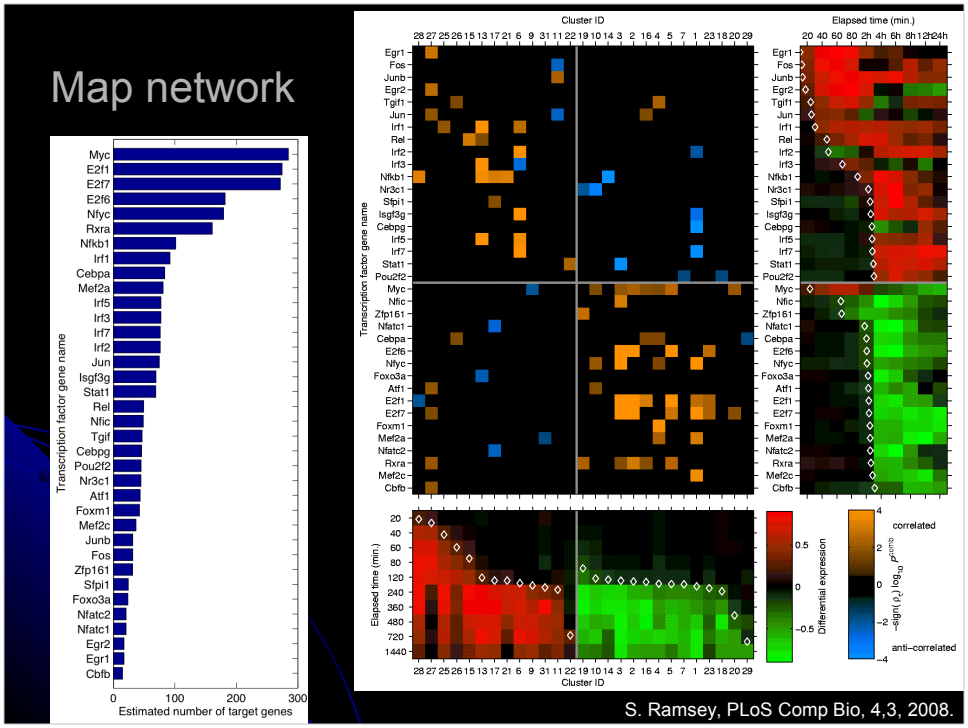
(5) probability of conservation based on genome alignments of 17 species (and a continuous-time Markov model for nucleotide substitutions and a two state phylo-HMM model to compute posterior conservation probabilities)

(6) Regulatory potential scores (log-likelihoods) that also make use of multiple genome alignments. After appropriate dimension reduction/alphabet selection, the method applies two variable order Markov models to estimate likelihoods of regulatory and neutral sites.

(7) The likelihood of binding to a non-functional binding site can be decreased by locating a stable nucleosome over those genomic regions while keeping functional sites accessible for TFs. Likelihood of nucleosome occupancies are computed using a method that uses a Markov model whose parameters are estimated from a set of known/measured nucleosome locations.

Data fusion can be performed in a Bayesian setting:

- We assume that *S* and *D* are conditionally independent and that the probability of *D* does not depend on the motif and background models.
- Conceptually, the probability *P* (*D* | *A; pi*) can be viewed as a positional prior for binding sites.

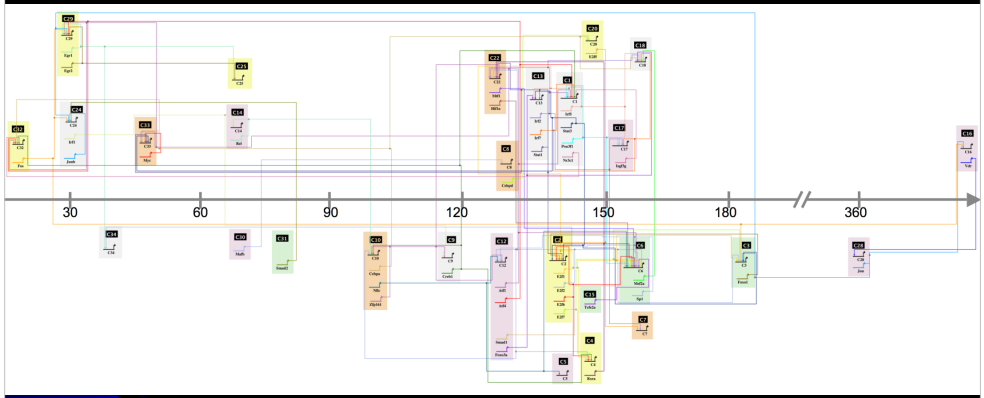


(LEFT) The histogram on the left shows the estimated “out degree” of each TF gene in the network. This is just a very rough estimate based on motif scanning data, but it shows a long-tailed distribution.

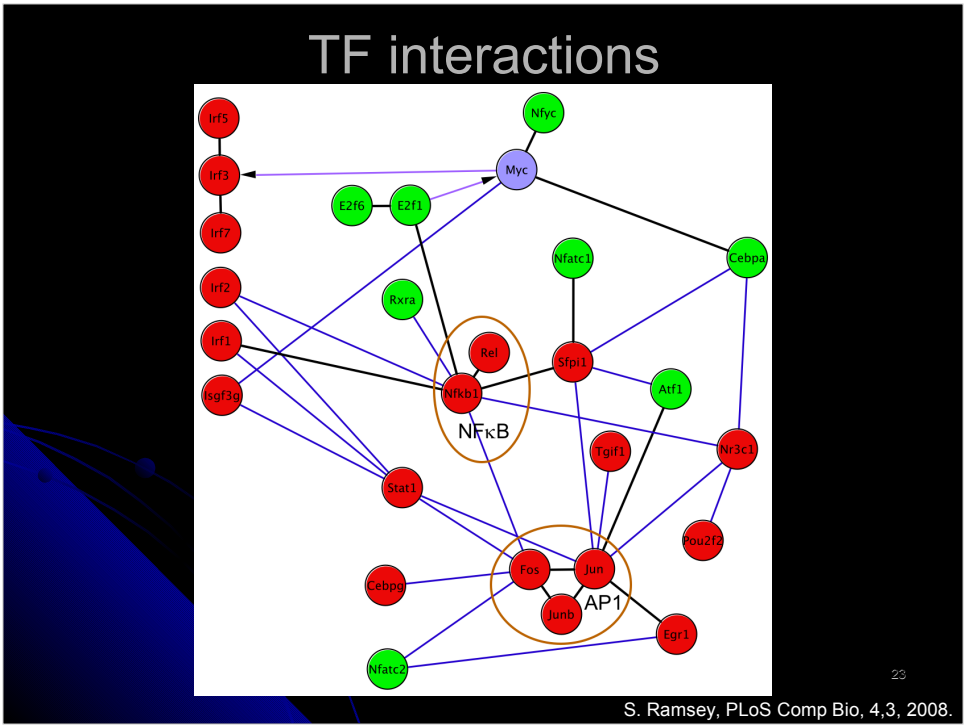
(RIGHT)

Transcription factor genes associated with clusters in the inferred transcriptional network. (A) The matrix shows associations between transcription factor genes and co-expressed gene clusters. Each column represents one of the 27 clusters within the inferred network, and each row represents one of the 36 transcription factor genes in the network. Clusters are ordered based on the LPS response time, defined as the time (under LPS stimulation) at which the cluster-median differential expression level reaches 25% of the maximum differential expression (see Materials and Methods, Expression Clustering). Transcription factor genes are ordered based on the LPS response time. The vertical gray line separates upregulated clusters (left half) from downregulated clusters (right half). The horizontal gray line separates upregulated transcription factors (top) from downregulated transcription factors (bottom). An orange or blue square indicates a statistically significant association between the transcription factor gene and the cluster, based on both promoter scanning and expression dynamics. An orange solid rectangle represents a positive average time-lagged correlation with genes in the cluster; a blue solid rectangle represents a negative average time-lagged correlation. (B) The red- green matrix is a heat-map showing transcription factor gene expression. The color indicates the normalized differential expression of the indicated ranscription factor gene (over time), in LPS-stimulated wild-type macrophages (SDR, see Equation 1). Red indicates upregulation relative to unstimulated macrophages and green indicates downregulation. A diamond symbol indicates the transcription factor response time. (C) Each column of the red-green matrix indicates the median normalized differential expression of the genes in the indicated cluster (over time), in LPS-stimulated wild-type macrophages. The diamond indicates the average LPS response time of the genes within the cluster.

TF-to-cluster association network

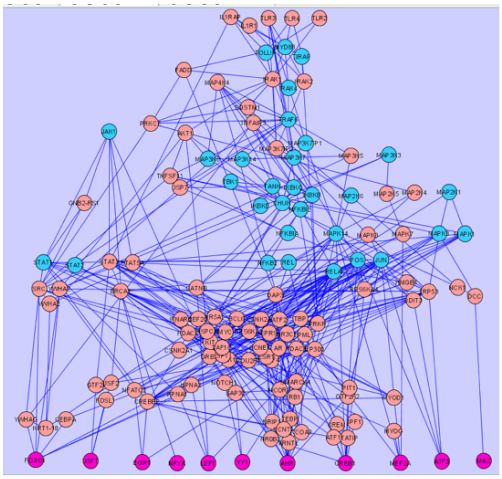


S. Ramsey, PLoS Comp Bio, 4,3, 2008.



Here, transcription factor genes are arranged in a cytoscape diagram in which each node is a TF gene, and each edge represents a protein-protein interaction (black or blue) or a protein-DNA interaction (purple arrow). Solid black edges further indicate that the two nodes are co-associated with at least one cluster. The color of the node indicates its differential expression under LPS stimulation (up or down).

TLR signaling pathway and downstream predicted transcription factors: using protein-protein interactions



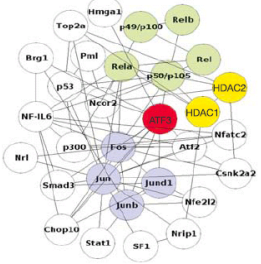
Known TLR Signaling pathway

(Known TFs)

Link to Upstream prot.

Link to Known TFs

Predicted TFs



ATF3 (red) is predicted to interact with a number of TFs, including members of the AP1 and NF-kB TF complexes.

B. Li, V. Thorsson

ChIP-on-chip validation

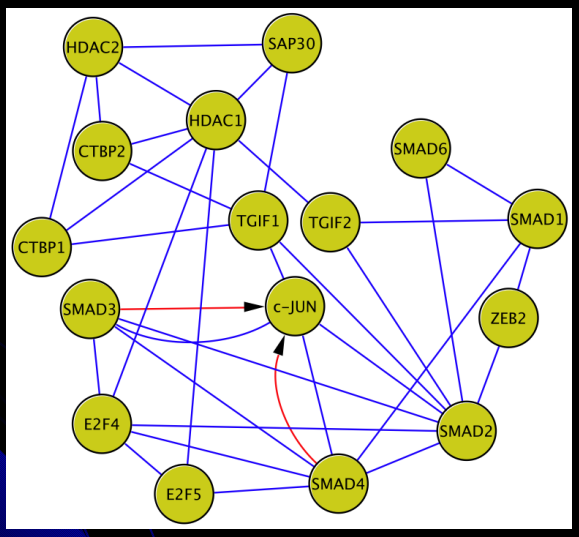
Table 2. Validation of transcription factor-to-cluster associations using ChIP-on-chip

TF	Matrix	Stim.	Clust	Time Points	In Clust	On Chip	Bound	*P*-Value
NFκB/p50	NFKB_Q6	LPS	C13	1 h, 2 h	64	23	18	1.1×10^{-3}
NFκB/p50	NFKB_Q6	LPS	C17	1 h, 2 h	58	20	11	2.5×10^{-1}
NFκB/p50	NFKAPPAB_01	LPS	C28	1 h, 2 h	28	21	20	1.1×10^{-6}
IRF1	IRF_Q6_01	LPS	C13	1 h, 2 h, 4 h	64	23	18	2.3×10^{-3}
IRF1	IRF_Q6_01	LPS	C25	1 h, 2 h, 4 h	37	22	18	8.8×10^{-4}

S. Ramsey, PLoS Comp Bio, 4,3, 2008.

We performed targeted validation of certain predicted (TF,cluster) transcriptional regulatory associations using ChIP-on-chip.

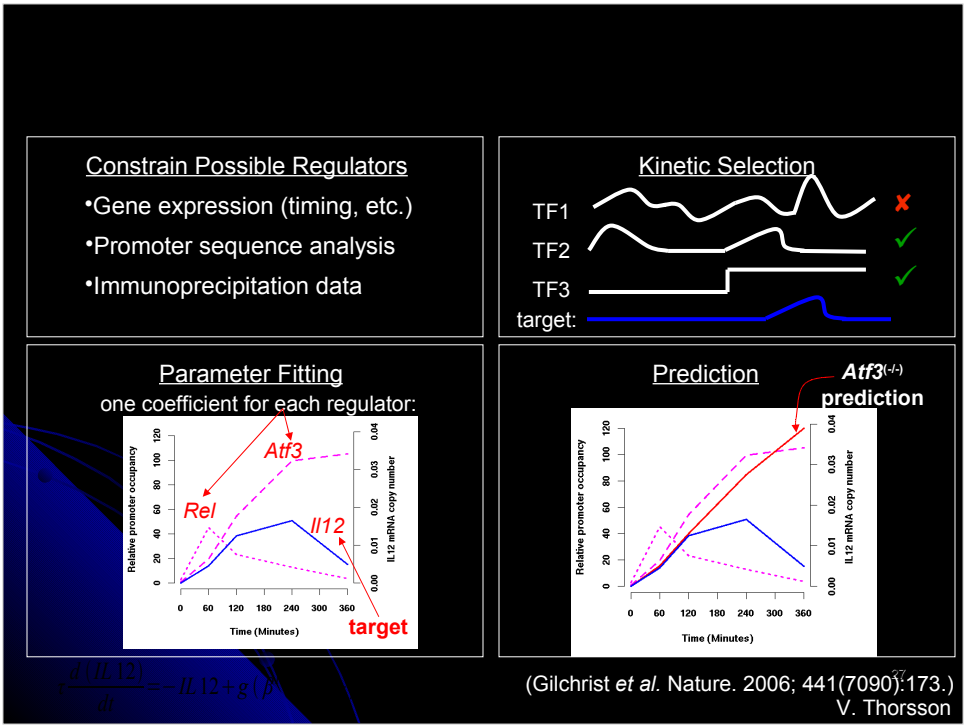
TGIF enhanceosome



28

S. Ramsey, PLoS Comp Bio, 4,3, 2008.

One transcription factor, TGIF1, was unexpected (it is not associated in the literature, with macrophage activation). We show here the network of TFs, co-factors, and other chromatin-modifying proteins with which TGIF1 associates in the protein interaction network. Blue edges indicate protein-protein interactions (HPRD), and red arrows indicate protein-DNA interactions (BOND).



Moving up the scale in terms of the complexity of the model class, we have Inferelator-type methods, by which I mean multivariate regression with an ODE-based model class. The basic idea is to try and use the combined time-courses of multiple TFs to try and best capture the derivative of the target gene's expression level, as estimated from time-course expression data. Because there are so many possible combinations of multiple TFs that could be used as predictors of a gene, it is in practice necessary to constrain the possible combinations of regulators using one or more sources of additional evidence. The choice of evidences will depend on the application, but generally include promoter sequence analysis and possibly ChIP data. The top-right shows a cartoon illustration of the method. Given the target gene expression profile shown in blue, only TF2 and TF3 could reasonbly serve as predictors of the targe gene's expression kinetics. Quantitatively, for each regulator within a proposed combination of regulators for a target gene, a fit coefficient is used. Multivariate regression is used to obtain the best-fit coefficients for the given regulator combination. In the example shown in the lower left, the two possible regulators are Rel and Atf3, and the target expression is the gene IL12. In the lower right, coefficients have been obtained for both regulators, and the red curve shows the predicted time-course expression for IL12 under an in-silico knockout of the TF ATF3. (SWITCH) The method was applied with great success to predict the expression levels of biclusters of genes in Halobacterium across hundreds of different experimental conditions. (SWITCH) The method is more recently being applied to the problem of predicting TF-to-target gene associations in the TLR-stimulated macrophage.

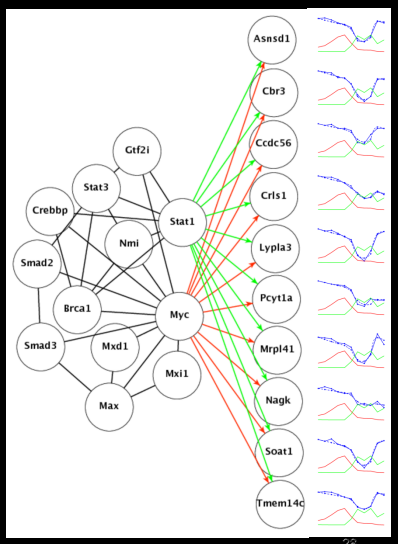
Network Predictions

Macrophage

Predicted network contains co-regulated gene groups with shared:

- Binding sites predictions, for single motifs or pairs of motifs
- Kinetics
- Predicted activators and repressors
- Predicted co-factors
- Function (GO categories)

Predictions can be tested with directed binding and/or functional assays

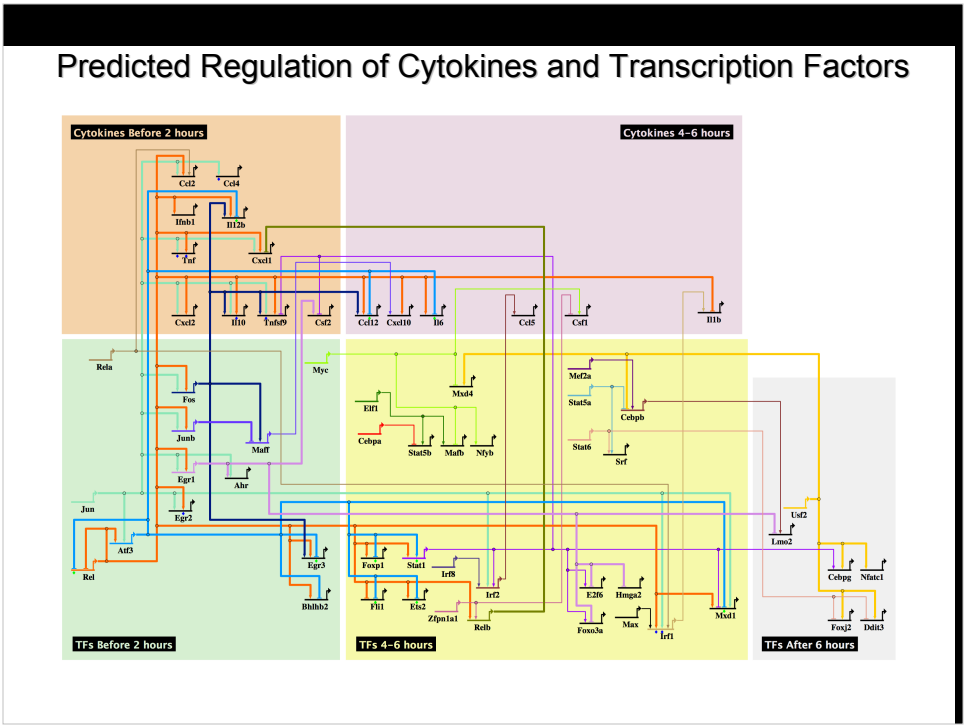


28

This is an example of a result of running the algorithm on our timecourse for LPS activation of BMDMS.

It shows of a panel of target genes (right) that are predicted to be under the control of a transcriptional complex involving Myc and Stat1.

Black edges, protein-protein interactions (from databases). Directed edges, predicted binding and regulation.Thumbnails illustrate target profiles (blue, predicted and measured, LPS timecourse) and Myc Profile, red, Stat1 profile green.



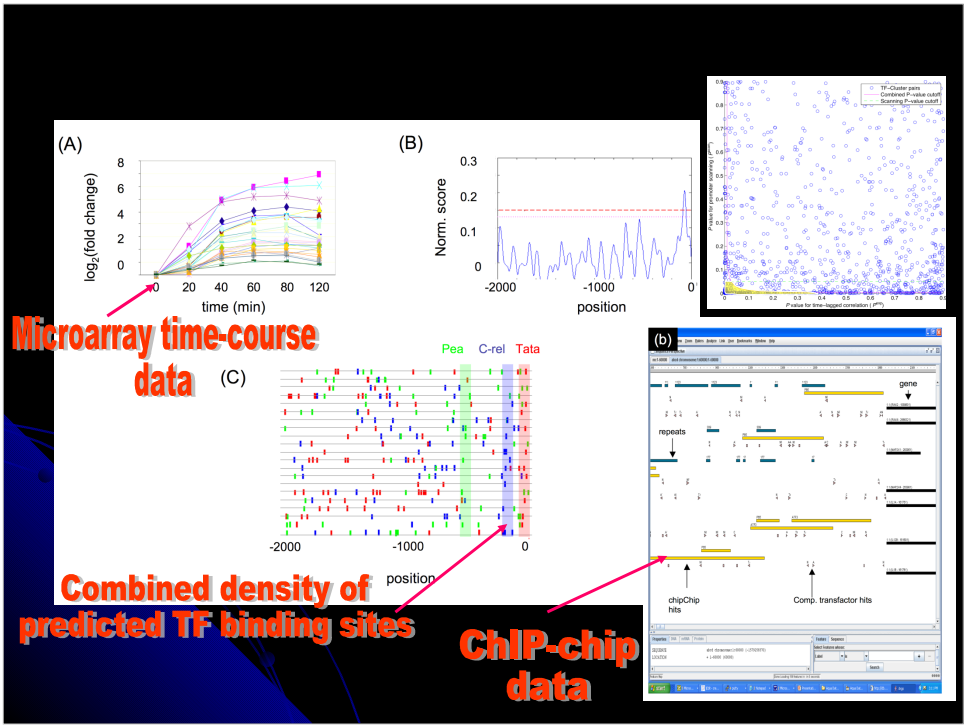
We are specifically interested in predictions for cytokines and for regulation of transcriptional regulators.

This figure shows network model predections for these functional groups as a BioTapestry network “wiring diagram”

The upper diagrams show possible regulation of cytokines over time, and the lower one possible regulation of TFs over time. The interdependency of the TFs implies a possible “cascade” of regualtion.

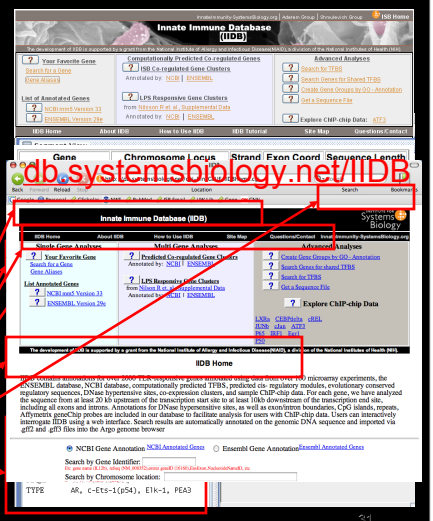
To test individual predictions, we perform directed experiments (coming slides).

(Explanatory note: The above diagram may not have been updated to include the predictions listed on the coming slides. CCL12 regulation *is* there however and could be used as a transition to the next slide)


(A)
log2(fold change)
time (min)
(B)
Norm. score
position
Microarray time-course data
(C)
Pea
C-rel
Tata
position
Combined density of predicted TF binding sites
(b)
gene
repeats
chipChip hits
Comp. transfactor hits
ChIP-chip data

Innate Immune Database (IIDB)

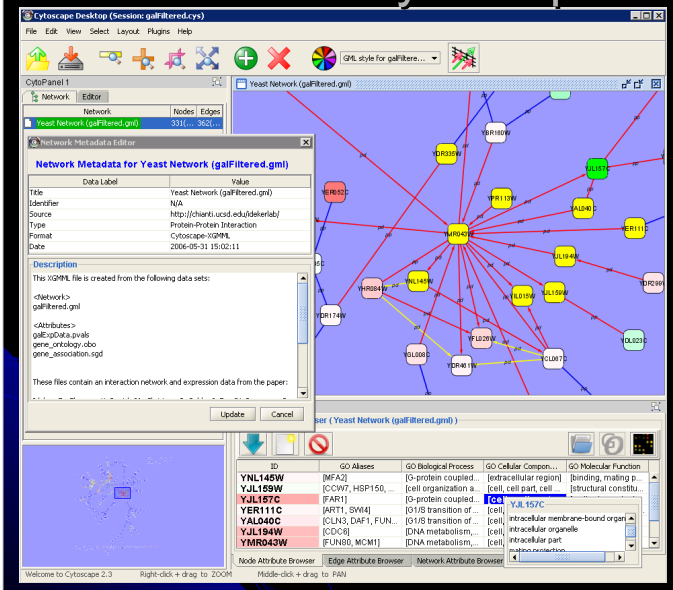
- Genomic annotations and *cis*-regulatory element predictions for immune-related genes.
- Web-based software tool for querying and visualization.
- Display expression time-course data (and clusters)
- Graphical visualization of genomic annotations combining many different data types



Korb *et al*. *BMC Immunology*, 9:7, 2008.

The innate immunity database is a web-based resource for viewing genome annotations for immunologically important genes. To date, the system includes around 2000 genes identified through differential expression in murine macrophages or RAW cells under stimulation by purified TLR agonists, or by a web-based request system for genes to be included in the system. In particular the system includes all genes represented on the custom Affymetrix tiling array that we are using for ChIP-on-chip experiments. For each gene, the system can display a large number of different genome annotation features graphically, including sequence conservation, computationally predicted transcription factor binding sites (and clusters thereof), and chromosomal segments identified through ChIP-on-chip experiments. The system can also graphically display the microarray-derived time-course expression data for a gene, under stimulation by various TLR agonists. Finally, the system allows the user to filter for genes that have evidence of binding by one or more selected transcription factors, based on computational TFBS motif scanning and/or ChIP-on-chip data. A publicly accessible version of this system is expected to be released to the scientific community at the time of publication. Here is the URL for anyone who is interested in trying it out.

Cytoscape



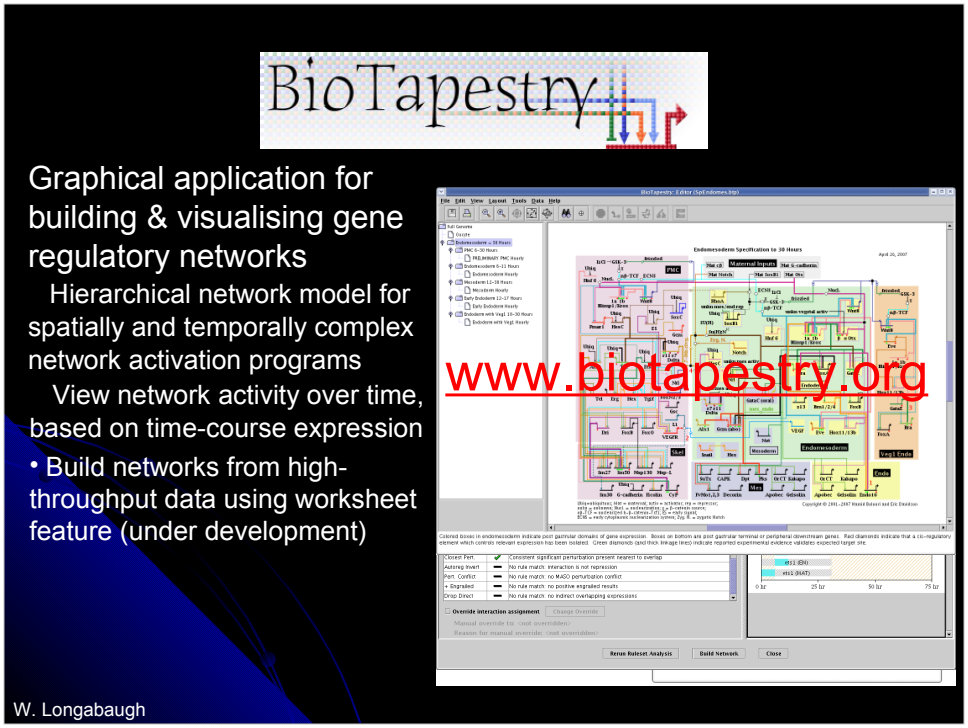
Visually Integrate gene expression, protein state, protein interactions, and protein class (ontology)

Analysis plug-in modules

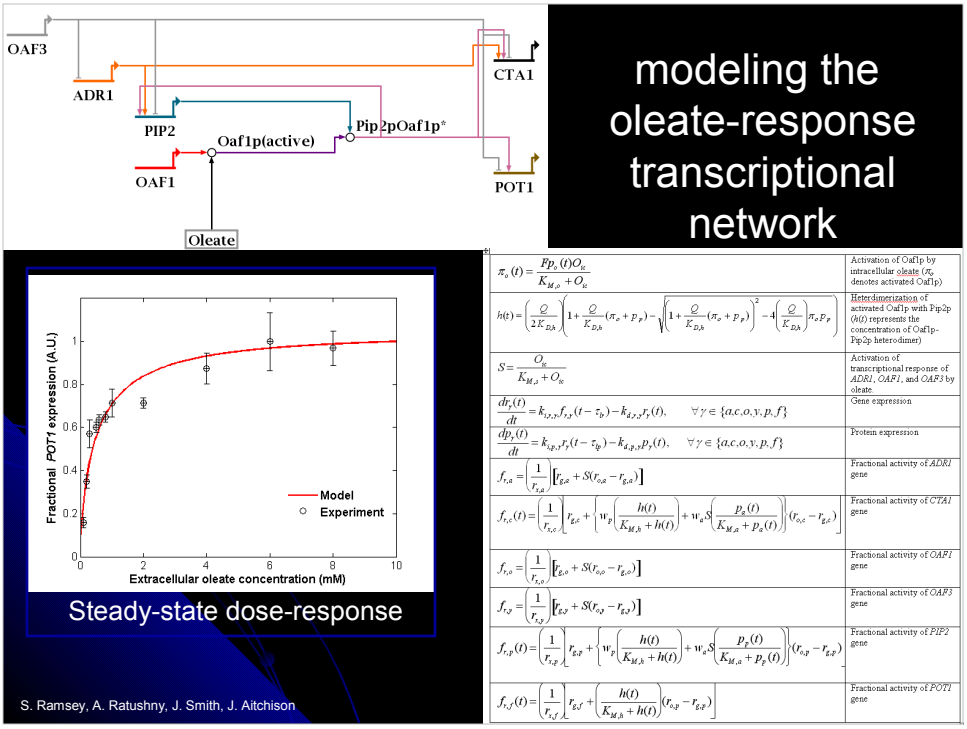
Implemented in Java

(networks, attributes, network metadata, etc.)

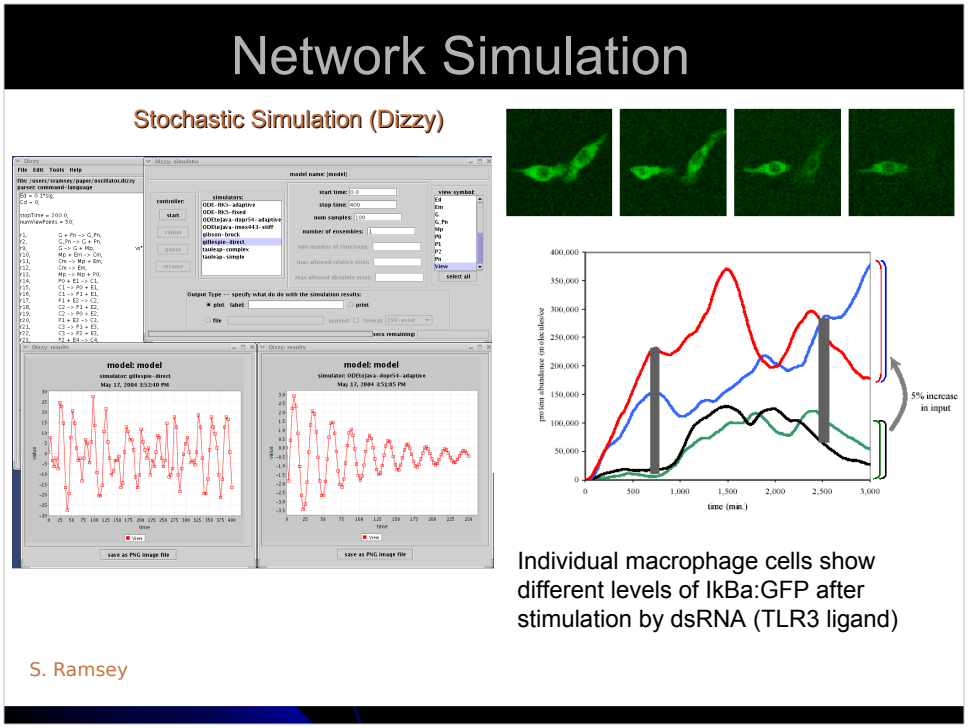
http://www.cytoscape.org/



BioTapestry is a software application for visualizing and buiding a gene regulatory network. It is specifiically adapted for networks that have spatially and temporally complex activation, such as in embryonic development. A key feature of BioTapestry is its ability to manage a hierarchy of sub-models, each of which could represent the portion of the network active within a different tissue or compartment of an organism, as well as to store information about the timing of activation of each gene within each sub-model. I would like to mention two recent developments with BioTapestry. First, BioTapestry has a new network building screen that enables the automated inference of a putative network from a large-scale perturbation expression dataset. This feature makes use of various heuristics and acts as an “expert system” in the interpretation of large-scale qPCR and WMISH experiments. Second, BioTapestry’s capabilities for large-scale network layout have been significantly improved, with a new layout style introducted specifically for transcription factor-to-cluster networks resulting from analysis using Inferelator-type methods.



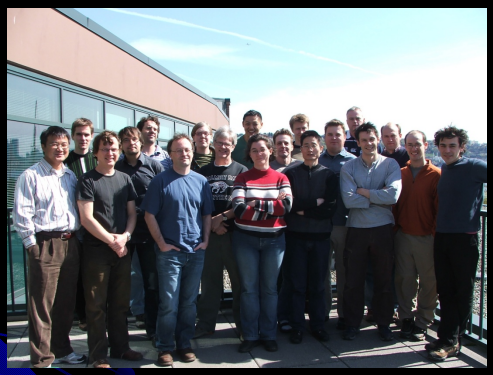
Moving up the scale of mechanistic detail a bit, one can build dynamic models in terms of biochemical kinetics (or more truthfully, so-called effective kinetic models, usually aggregating multiple interactions into a single effective reaction). We are using such an approach to model the dynamics of the regulatory network controlling the response to oleic acid in yeast, shown in cartoon form on the upper right. For this system, we have access to three types of data that we use to constrain and guide model development: steady-state dose response of a key reporter gene activity (lower left), to the concentration of extracellular oleate; time-course expression of key TFs and target genes under a carbon source transition from glycerol to oleate (both using qPCR and microarrays); and oleate-to-glycerol relative gene expression levels for deletion strains of key TFs. In addition, there is quite a bit of useful data on gene activities of key targe genes for yeast at steady-state growth conditions on various carbon sources. Our broad goal in developing this kinetic model is to be able to gain insight into how the architecture (i.e., wiring diagram and positive/negative lines of control) of the transcription network was selected for, and what dynamical properties of the oleate response are resultant from specific elements of this architecture. Oaf3p is a particular case in point, where we wish to explore whether Oaf3p acts as a fast-responding buffer to prevent transient de-activation of the genetic switch in the event of a drop in the concentration of intracellular oleate. Additionally, we wish to explore whether heterodimerization of two TFs acts as a nose filter (lower right).



(a)-(d) Macrophage heterogeneity in signaling. Video microscopy of a clone of two individual RAW 264.7 macrophage cells expressing human-IκBα-EGFP during stimulation with 1 microgram/ml double-stranded-RNA (a ligand that binds Toll-like receptor 3). The decline in EGFP fluorescence (cell marked by red arrow) is indicative of IκBα degradation during signaling leading to NF-κB activation. **(e) Quantification of IκBα degradation in two individual cells.** The fluorescence of each cell in **(a)**-**(d)** was quantified and normalized to their initial values. Note how IκBα degradation starts at the same rate in both cells, then at about 40 minutes the cell on the right suddenly degrades its IκBα much faster and further than the cell on the left**.** The delayed onset of differential behavior is consistent with the hypothesis that the difference between the two cells may be due to transcriptional noise in genes whose products interact with the signal transduction pathway upstream of IκBα.
0 min, 28 min, 44 min, 69 min
****

Simulated stochastic transcriptional noise in macrophages. A simulation of four protein levels in macrophage. The green and black protein levels are driven by the same transcription factor, and the red and blue protein levels are driven by another transcription factor that has an elevated (+5%) average abundance. The variations within each curves and also between the blue and red protein levels, and between the green and black protein levels, is due to the intrinsic noise in transcription. In contrast to *E. coli* and yeast genes, which exhibit rapidly changing (spiky) protein levels, in macrophages intrinsic variations in protein levels occur very slowly (due to much slower mRNA and protein degradation rates), on a timescale of hours. This scenario mimics two cells with 5% random difference in their cellular content and illustrates how this small difference, when amplified by slow intrinsic variations in gene expression can result in cells that are highly heterogeneous in terms of cellular content over periods of the order of 20 hours.

Acknowledgments



Steve Ramsey
Vesteinn Thorsson
Alistair Rust
Bin Li
John Boyle
Sandy Klemm
Harri Lähdesmäki
Martin Korb
Matti Nykter
Sarah Killcoyne
Chris Cavnor
Bill Longabaugh
Antti Niemistö
Ricardo Vencio

Aderem Group

Support

NIH/NIGMS R21 GM070600
NIH/NIGMS R01 GM072855
NIH/NIAID U54 AI54253
NIH/NIGMS P50 MO-76547