Bioinformatics 3

V8 – Gene Regulation

Mon, Nov 23, 2015

- Measuring transcription + translation rates
- Motifs in GRN
- Reconstruction of GRNs – basic methods

Rates of mRNA transcription and protein translation

ARTICLE

doi:10.1038/nature10098

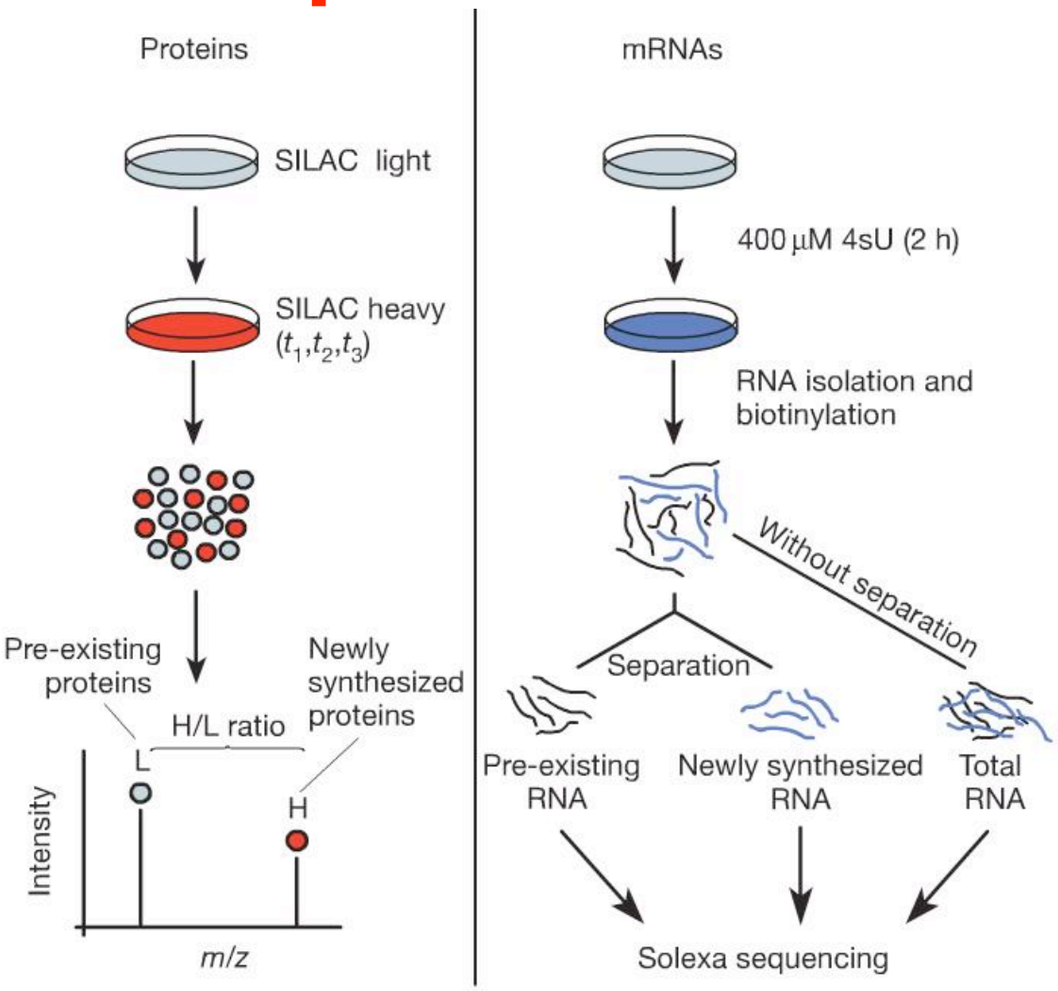
Global quantification of mammalian gene expression control

Björn Schwanhäusser[1], Dorothea Busse[1], Na Li[1], Gunnar Dittmar[1], Johannes Schuchhardt[2], Jana Wolf[1], Wei Chen[1] & Matthias Selbach[1]

SILAC: „stable isotope labelling by amino acids in cell culture“ means that cells are cultivated in a medium containing **heavy** stable-isotope versions of **essential amino acids**.

When non-labelled (i.e. light) cells are transferred to heavy SILAC growth medium, newly synthesized proteins incorporate the heavy label while pre-existing proteins remain in the light form.

Schwanhäuser et al. Nature 473, 337 (2011)



Parallel quantification of mRNA and protein turnover and levels. Mouse fibroblasts were pulse-labelled with heavy amino acids (SILAC, left) and the nucleoside **4-thiouridine** (4sU, right).
Protein and mRNA turnover is quantified by mass spectrometry and next-generation sequencing, respectively.

Rates of mRNA transcription and protein translation

84,676 peptide sequences were identified by MS and assigned to 6,445 unique proteins.

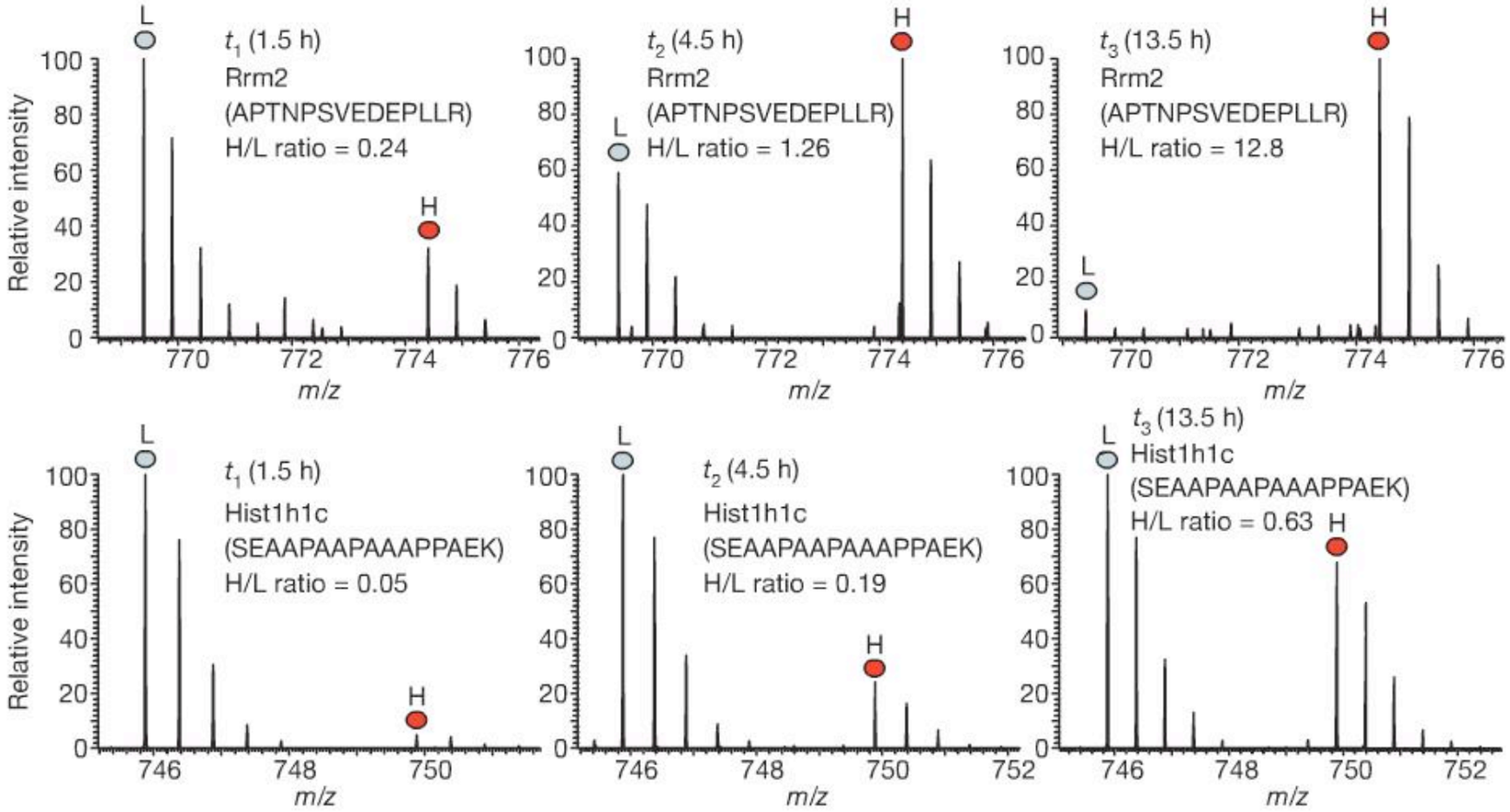
5,279 of these proteins were quantified by at least three heavy to light (H/L) peptide ratios belonging to these proteins.

Mass spectra of peptides for two proteins (x-axis: mass over charge ratio).

Top: high-turnover protein
Bottom: low-turnover protein.

Over time, the heavy to light (H/L) ratios increase.

You should understand these spectra!



Schwanhäuser et al. Nature 473, 337 (2011)

Protein half-lifes and decay rates

Extract ratio r of protein with heavy amino acids (P_H) and light amino acids (P_L):

$$r = \frac{P_H}{P_L}$$

Assume that proteins labelled with light amino acids decay exponentially with degradation rate constant k_{dp} :

$$P_L = P_0 e^{-k_{dp} t} .$$

Express (P_H) as difference between total number of a specific protein P_{total} and P_L:

$$P_H(t) = P_{total}(t) - P_L(t)$$

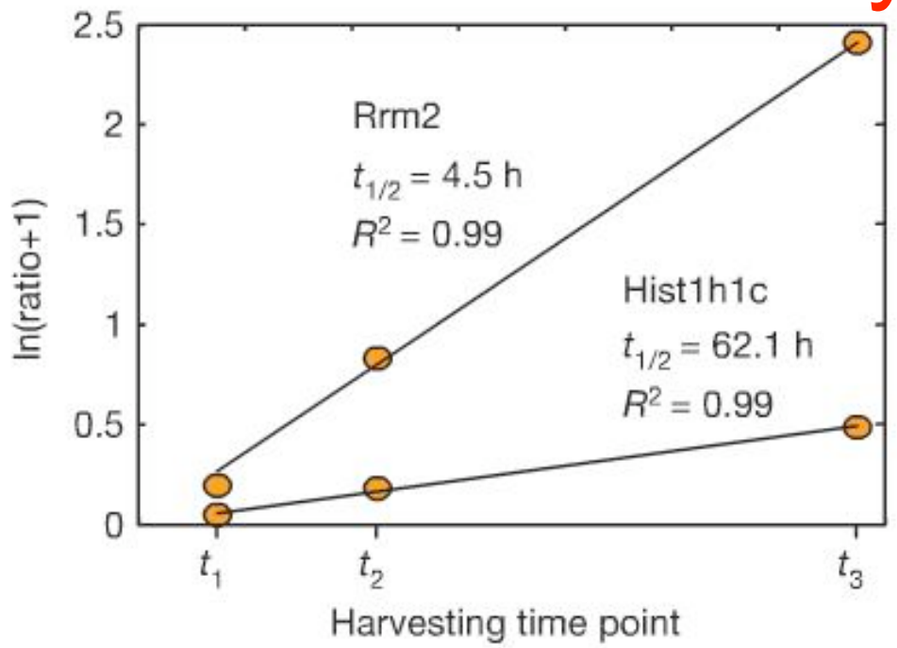
Assume that P_{total} doubles during duration of one cell cycle (which lasts t_∞):

$$P_H(t) = P_{total}(t) - P_L(t) = P_0 2^{t/t_{cc}} - P_L(t),$$

$$r = \frac{P_H}{P_L} = \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}} - 1$$

$$\frac{P_H}{P_L} + 1 = \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}}$$

$$\ln(ratio + 1) = \ln \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}} = \ln e^{k_{dp} t} + \ln 2^{\frac{t}{t_{cc}}} = k_{dp} t + \ln 2^{\frac{t}{t_{cc}}}$$



Consider m intermediate time points:

$$k_{dp} = \frac{\sum_{i=1}^{m} \log_e (r_{t_i} + 1) t_i}{\sum_{i=1}^{m} t_i^2} - \frac{\log_e 2}{t_{cc}},$$

gives the desired quantity (half-life):

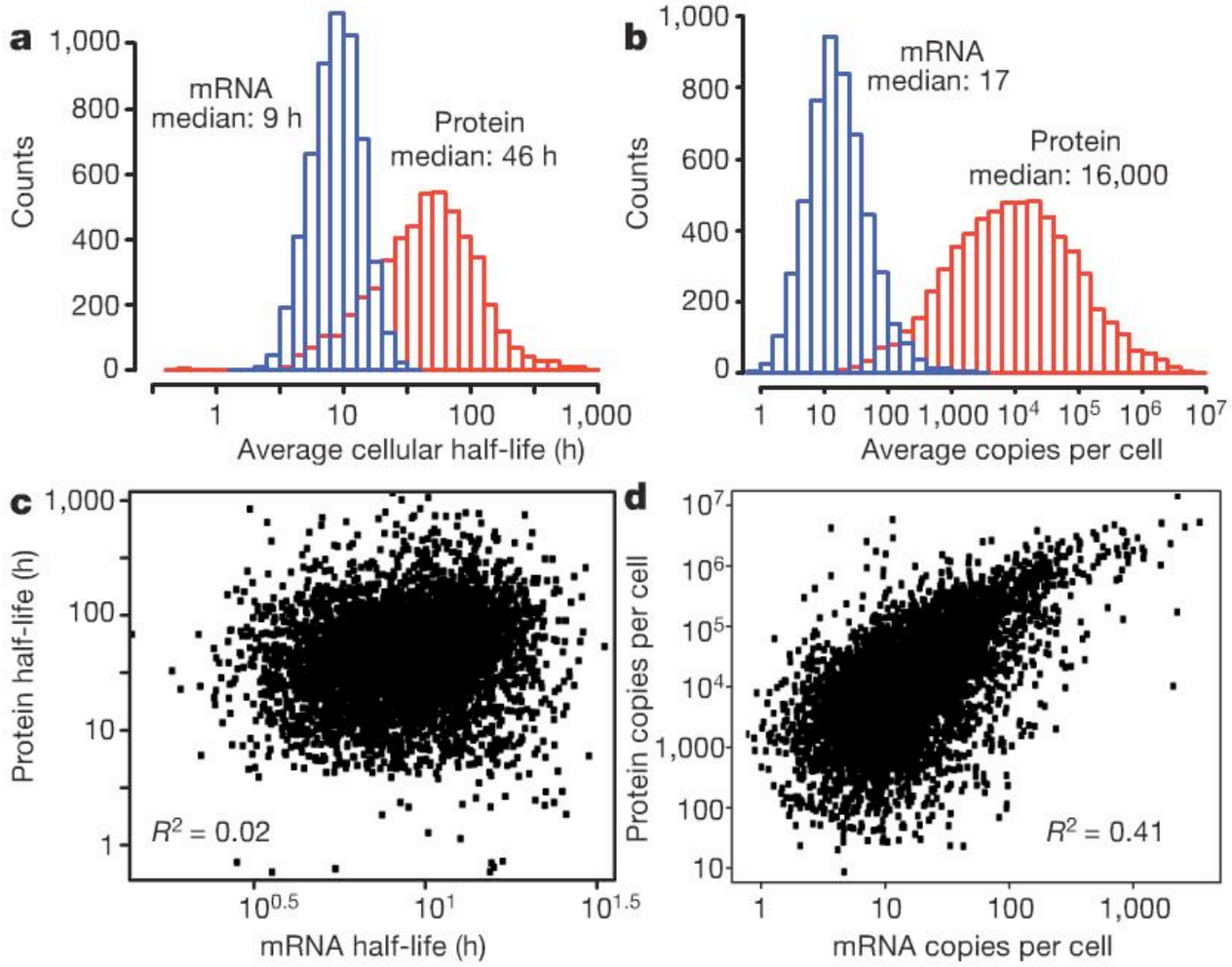
$$T_{1/2} = \frac{\log_e 2}{k_{dp}} .$$ because this gives

$$P_L = P_0 e^{-k_{dp} t} = P_0 e^{-k_{dp} \frac{\log_e 2}{k_{dp}}} = P_0 e^{\log_e \frac{1}{2}} = \frac{1}{2} P_0$$

The same is done to compute mRNA half-lives (not shown).

Schwanhäuser et al. Nature 473, 337 (2011)

mRNA and protein levels and half-lives



a, b, Histograms of mRNA (blue) and protein (red) half-lives (a) and levels (b).

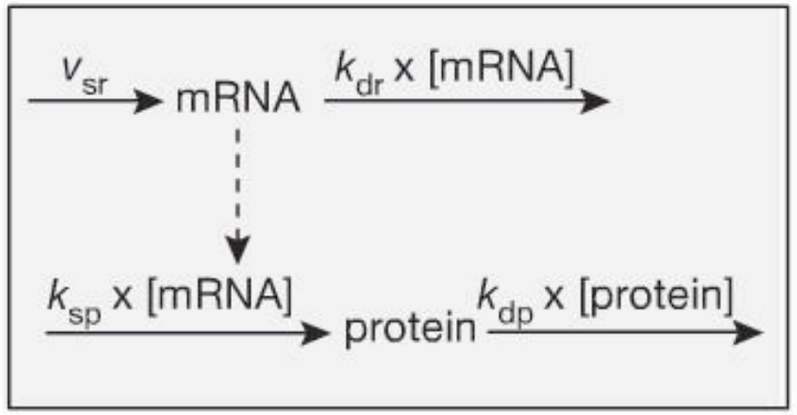
Proteins were on average 5 times more stable (9h vs. 46h) and 900 times more abundant than mRNAs and showed more variation.

c, d, Although mRNA and protein levels correlated significantly (right), correlation of protein and mRNA half-lives was virtually absent (left).

Schwanhäuser et al. Nature 473, 337 (2011)

Mathematical model of transcription and translation

A widely used minimal description of the dynamics of transcription and translation includes the synthesis and degradation of mRNA and protein, respectively



$$\frac{dR}{dt} = v_{sr} - k_{dr}R$$

$$\frac{dP}{dt} = k_{sp}R - k_{dp}P$$

The mRNA (R) is synthesized with a constant rate v_{sr} and degraded proportional to their numbers with rate constant k_{dr}.

The protein level (P) depends on the number of mRNAs, which are translated with rate constant k_{sp}.

Protein degradation is characterized by the rate constant k_{dp}.

The synthesis rates of mRNA and protein are calculated from their measured half lives and levels.

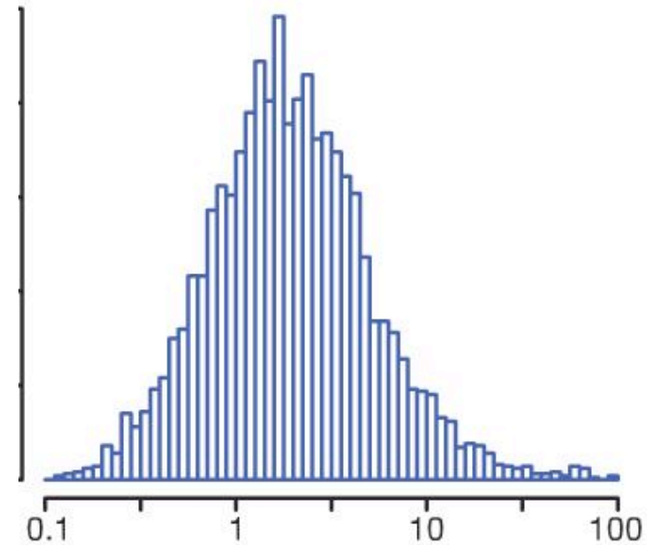
Schwanhäuser et al. Nature 473, 337 (2011)

Computed transcription and translation rates

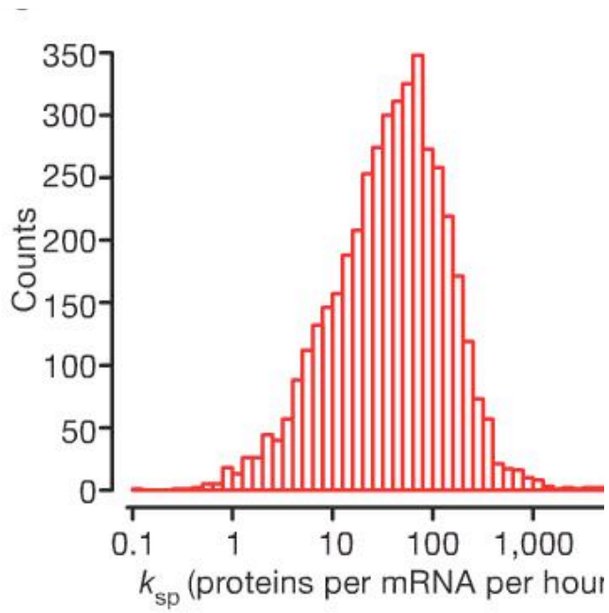
Average cellular transcription rates predicted by the model span two orders of magnitude.

The median is about 2 mRNA molecules per hour (**very slow**!).

An extreme example is the protein Mdm2 of which more than 500 mRNAs per hour are transcribed.



The median translation rate constant is about 40 proteins per mRNA per hour



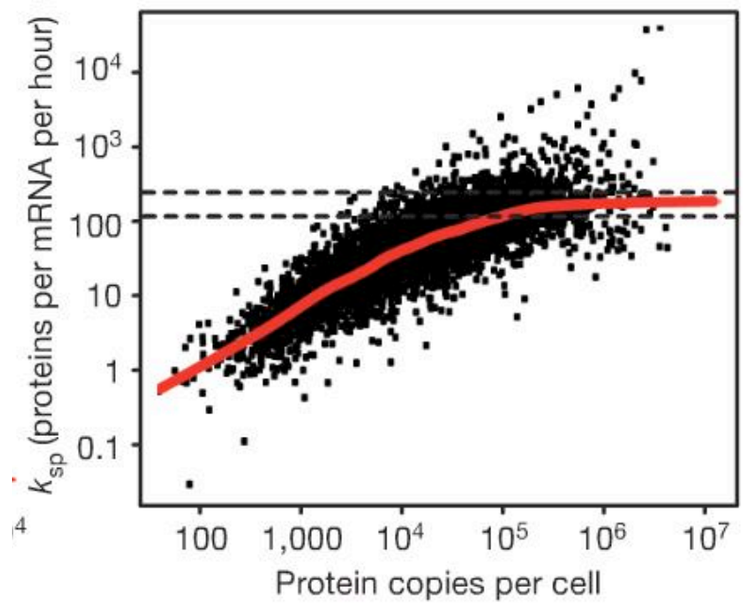
Calculated translation rate constants are not uniform

Schwanhäuser et al. Nature 473, 337 (2011)

Maximal translation constant

Abundant proteins are translated about 100 times more efficiently than those of low abundance

Translation rate constants of abundant proteins saturate between approximately 120 and 240 proteins per mRNA per hour.



The maximal translation rate constant in mammals is not known.

The estimated maximal translation rate constant in sea urchin embryos is 140 copies per mRNA per hour, which is surprisingly close to the prediction of this model.

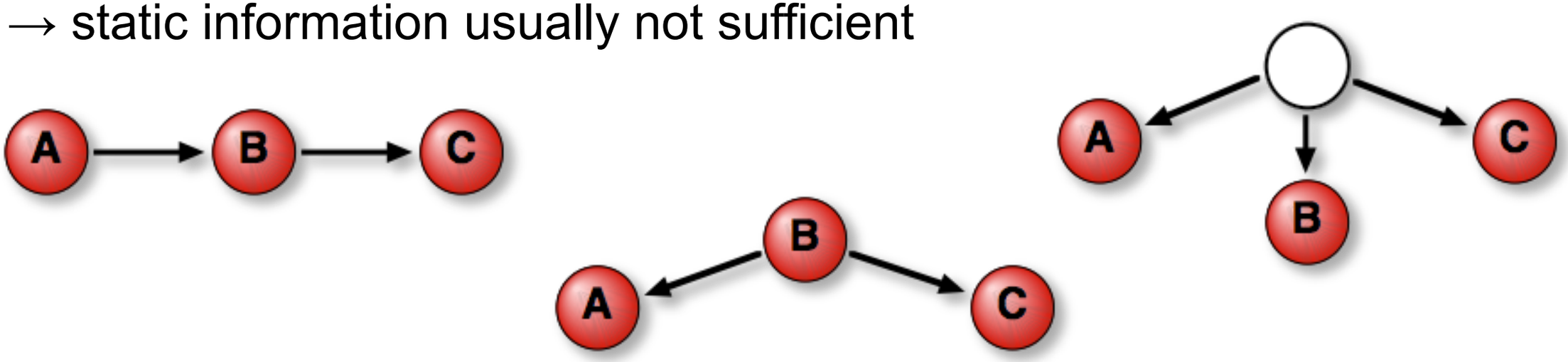
Schwanhäuser et al. Nature 473, 337 (2011)

Network Reconstruction

Experimental data: DNA microarray → expression profiles

Clustering → genes that are **regulated simultaneously**
→ Cause and action??? Are all genes known???

Shown below are 3 different networks that lead to the same expression profiles
→ **combinatorial explosion** of number of compatible networks
→ static information usually not sufficient



Some formalism may help
→ **Bayesian networks** (formalized conditional probabilities)
but usually too many candidates…

Network Motifs

Network motifs in the transcriptional regulation network of *Escherichia coli*

Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

Nature Genetics **31** (2002) 64

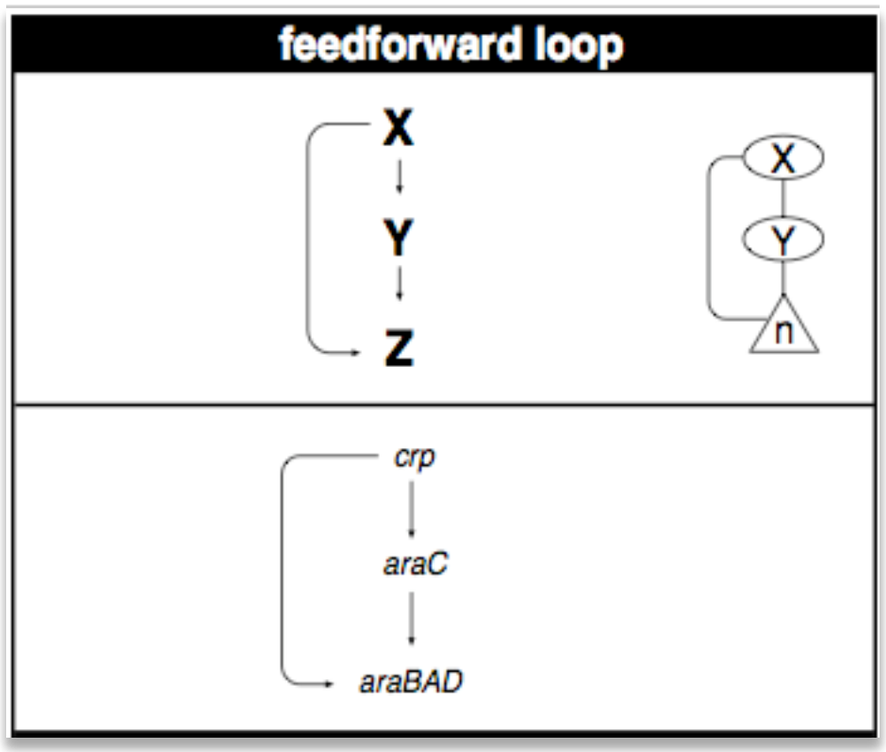
RegulonDB + their own hand-curated findings

→ break down network into motifs

→ statistical significance of the motifs?

→ behavior of the motifs <=> location in the network?

Motif 1: Feed-Forward-Loop



X = general transcription factor
Y = specific transcription factor
Z = effector operon(s)

Example for this in *E. coli*:
araBAD operon, encodes enzymes needed for the catabolism of arabinose

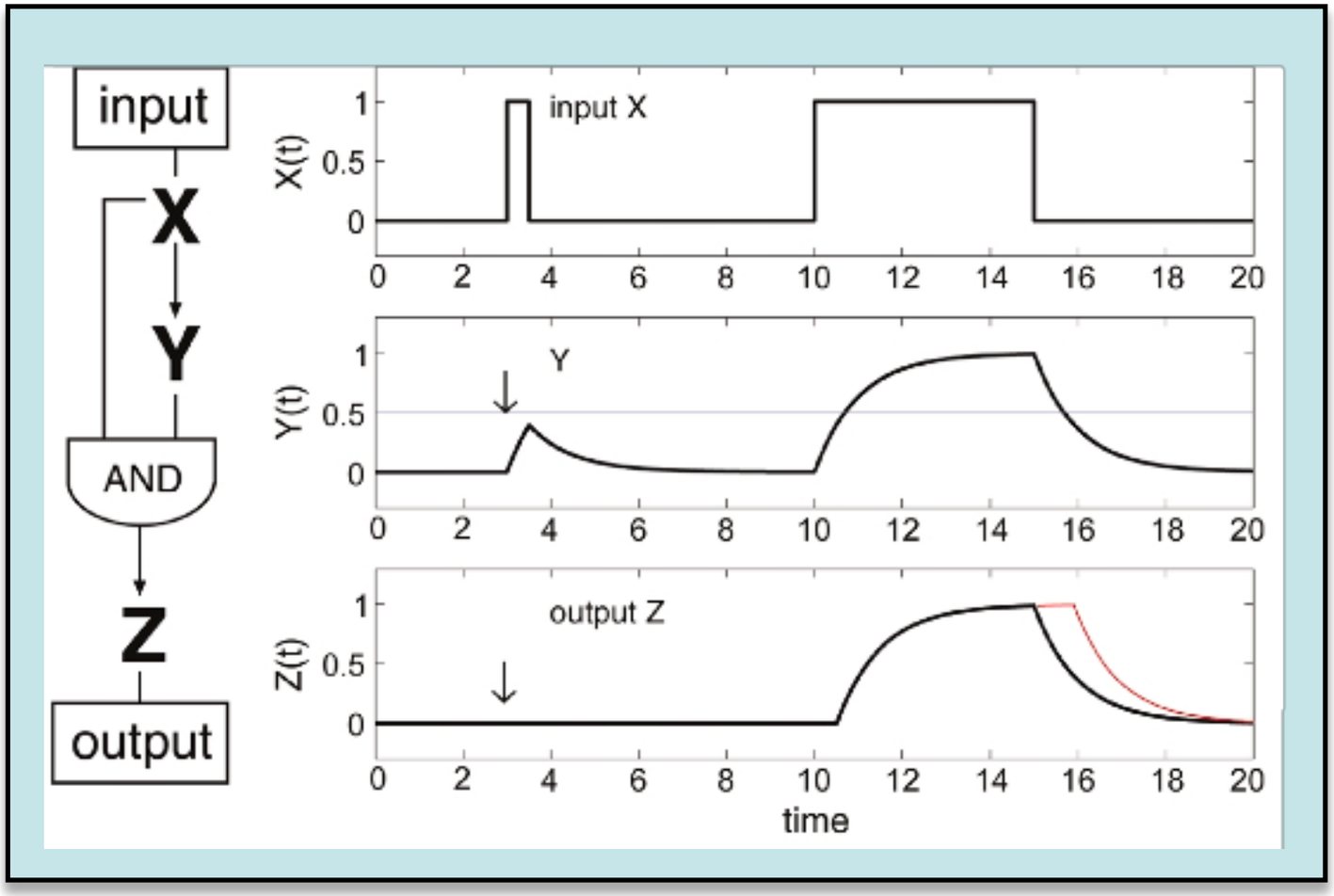
X and Y **together** regulate Z:

"**coherent**", if X and Y have the **same** effect on Z (activation vs. repression), otherwise "incoherent"

85% of the FFLs in *E. coli* are coherent

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

FFL dynamics



In a coherent FFL:
X **and** Y activate Z

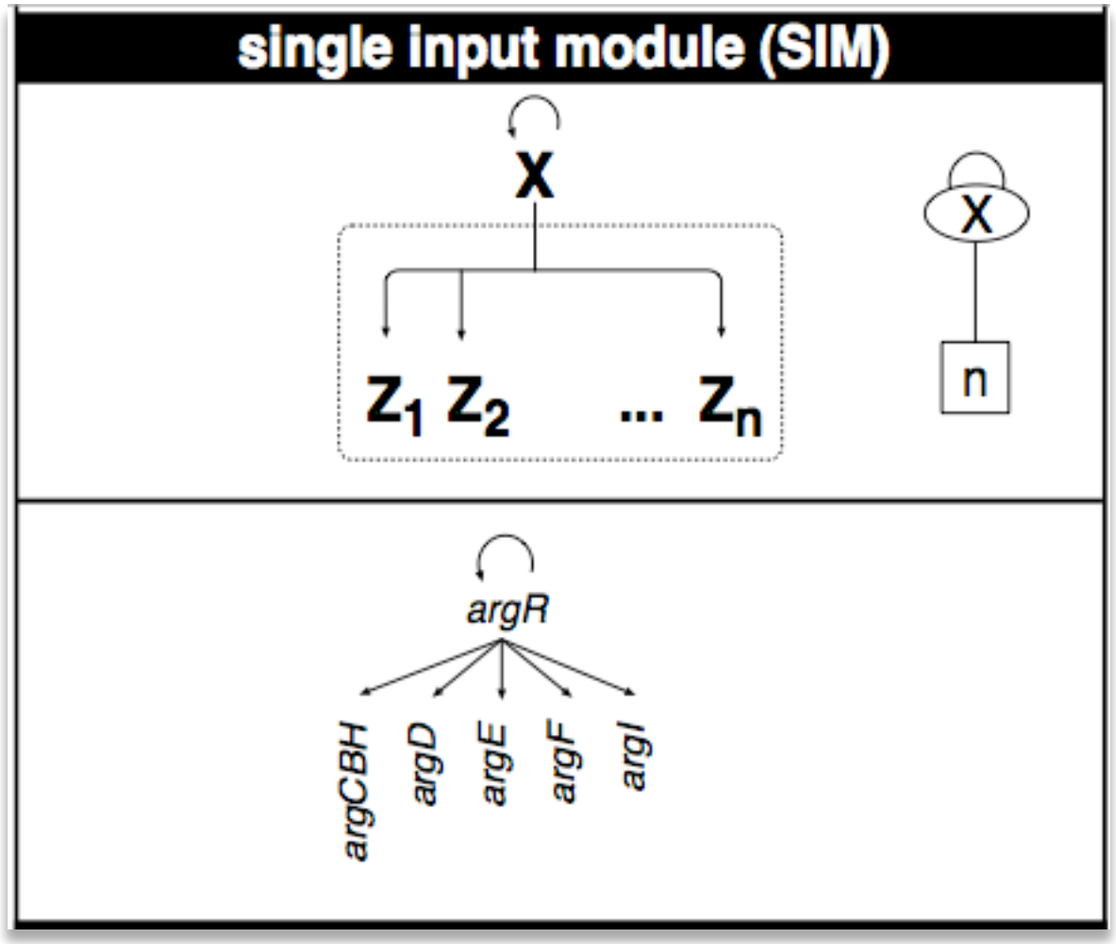
Dynamics:
- input activates X
- X activates Y (delay)
- (X && Y) activates Z

Delay between X and Y → signal must persist longer than delay
→ reject transient signal, react only to **persistent** signals
→ enables fast shutdown

Helps with **decisions** based on **fluctuating signals**

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Motif 2: Single-Input-Module



Set of operons controlled by a single transcription factor

- same sign
- no additional regulation
- control is usually autoregulatory (70% vs. 50% overall)

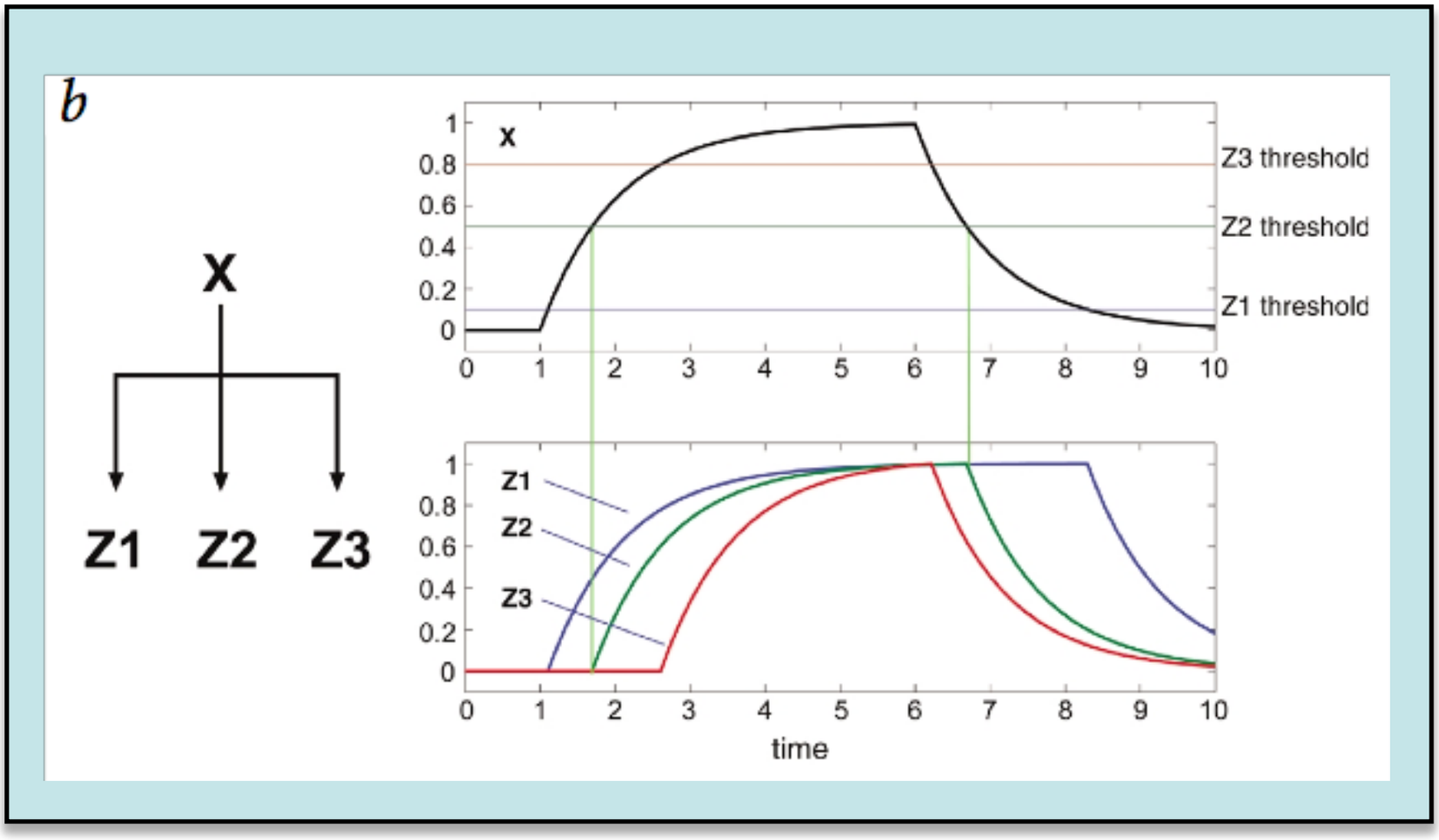
Example for this in *E. coli*: arginine biosynthetic operon *argCBH* plus other enzymes of arginine biosynthesis pathway

Mainly found in genes that code for **parts** of a protein **complex** or metabolic **pathway**

→ produces components in comparable amounts (stoichiometries)

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

SIM-Dynamics



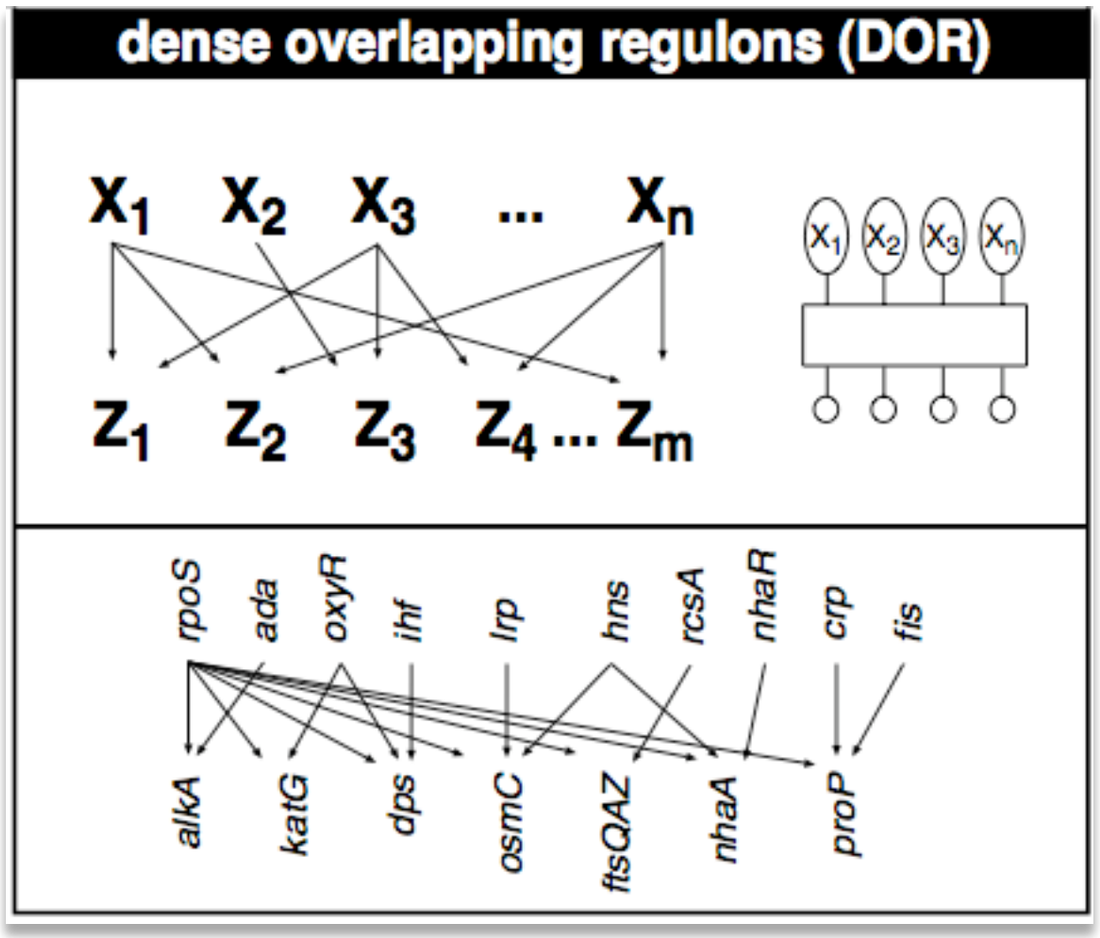
If different thresholds exist for each regulated operon:

→ first gene that is activated is the last that is deactivated

→ well defined temporal ordering (e.g. flagella synthesis) + stoichiometries

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Motif 3: Densely Overlapping Regulon



Dense layer between groups of transcription factors and operons
→ much denser than network average (≈ community)

Usually each operon is regulated by a different combination of TFs.

Main "**computational" units** of the regulation system

Sometimes: same set of TFs for group of operons → "multiple input module"

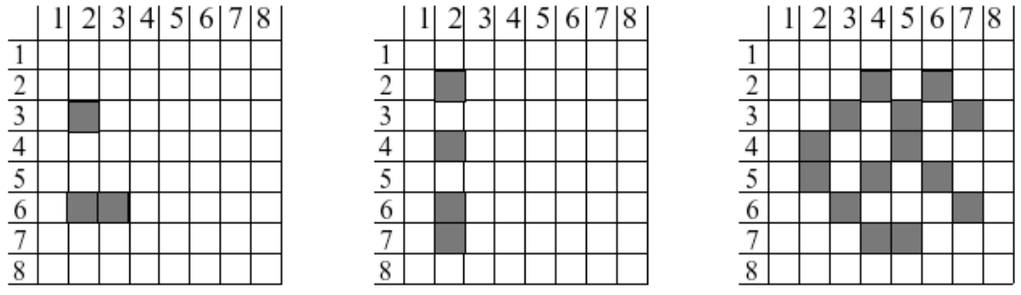
Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Detection of motifs

Represent transcriptional network as a connectivity matrix *M* such that M_{ij} = 1 if operon *j* encodes a TF that transcriptionally regulates operon *i*
and M_{ij} = 0 otherwise.

Scan all $n \times n$ submatrices of *M* generated by choosing *n* nodes that lie in a connected graph, for *n* = 3 and *n* = 4.

Submatrices were enumerated efficiently by recursively searching for nonzero elements.



Connectivity matrix for causal regulation of transcription factor *j* (row) by transcription factor *i* (column). Dark fields indicate regulation. (Left) Feed-forward loop motif. TF 2 regulates TFs 3 and 6, and TF 3 again regulates TF 6. (Middle) Single-input multiple-output motif. (Right) Densely-overlapping region.

For *n* = 3, the only significant motif is the feedforward loop.
For *n* = 4, only the overlapping regulation motif is significant.
SIMs and multi-input modules were identified by searching for identical rows of *M*.

Shen-Orr et al. Nature Gen. 31, 64 (2002)

Motif Statistics

Compute a p-value for submatrices representing each type of connected subgraph by comparing # of times they appear in real network vs. in random network.

Table 1 • Statistics of occurrence of various structures in the real and randomized networks

Structure	Appearances in real network	Appearances in randomized network (mean ± s.d.)	*P* value
Coherent feedforward loop	34	4.4 ± 3	$P < 0.001$
Incoherent feedforward loop	6	2.5 ± 2	$P \sim 0.03$
Operons controlled by SIM (>13 operons)	68	28 ± 7	$P < 0.01$
Pairs of operons regulated by same two transcription factors	203	57 ± 14	$P < 0.001$
Nodes that participate in cycles*	0	0.18 ± 0.6	$P \sim 0.8$

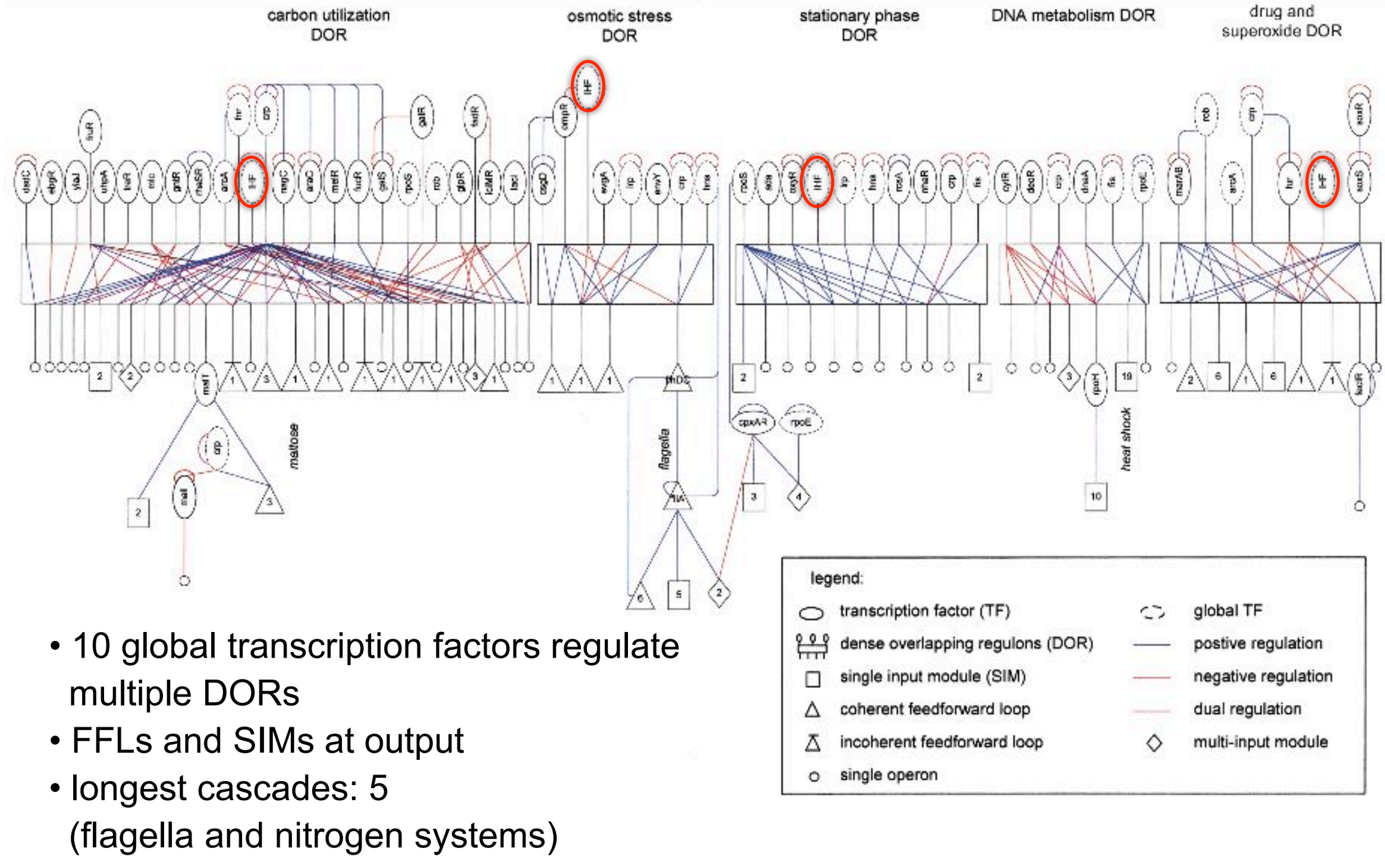
*Cycles include all loops greater than size 1 (autoregulation). *P* value for cycles is the probability of networks with no loops.

All motifs are highly **overrepresented** compared to randomized networks

No cycles ($X \rightarrow Y \rightarrow Z \rightarrow X$) were identified, but this was not statistically significant in comparison to to random networks

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Network with Motifs



- 10 global transcription factors regulate multiple DORs
- FFLs and SIMs at output
- longest cascades: 5 (flagella and nitrogen systems)

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

gene-regulatory networks

What are gene-regulatory networks (GRNs)?

- networks between genes coding for transcription factors and genes

How does one generate GRNs?

- from co-expression + regulatory information (e.g. presence of TF binding sites)

What can these GRNs be used for?

functional interpretation of exp. data, guide inhibitor design etc.

Limitations of current GRN models:

incomplete in terms of TF-interactions,
usually do not account for epigenetic effects and miRNAs

How does one generate GRNs?

(1.) „by hand“ based on individual experimental observations

(2) Infer GRNs by computational methods from gene expression data (see reference below)

Briefings in Bioinformatics Advance Access published May 21, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 17 doi:10.1093/bib/bbt034

Supervised, semi-supervised and unsupervised inference of gene regulatory networks

Stefan R. Maetschke, Piyush B. Madhamshettiwar, Melissa J. Davis and Mark A. Ragan

Submitted: 19th January 2013; Received (in revised form): 15th April 2013

Unsupervised methods

Unsupervised methods are either based on **correlation** or on **mutual information**.

Correlation-based network inference methods assume that correlated expression levels between two genes are indicative of a regulatory interaction.

Correlation coefficients range from -1 to 1.

A **positive** correlation coefficient indicates an **activating interaction**, whereas a **negative** coefficient indicates an **inhibitory interaction**.

The common correlation measure by **Pearson** is defined as

$$corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma(X_i) \cdot \sigma(X_j)}$$

where X_i and X_j are the expression levels of genes *i* and *j*, cov(.,.) denotes the covariance, and σ is the standard deviation.

Rank-based unsupervised methods

Pearson's correlation measure assumes normally distributed values.
This assumption does not necessarily hold for gene expression data.

Therefore rank-based measures are frequently used.
The measures by Spearman and Kendall are the most common.

Spearman's method is simply Pearson's correlation coefficient for the ranked expression values

Kendall's τ coefficient : $$\tau(X_i, X_j) = \frac{con(X_i^r, X_j^r) - dis(X_i^r, X_j^r)}{\frac{1}{2} n(n-1)}$$

where X^r_i and X^r_j are the ranked expression profiles of genes *i* and *j*.

Con(.) denotes the number of concordant value pairs (i.e. where the ranks for both elements agree). *dis(.)* is the number of disconcordant value pairs in X^r_i and X^r_j. Both profiles are of length *n*.

WGCNA

WGCNA is a modification of correlation-based inference methods that **amplifies high correlation coefficients** by raising the absolute value to the power of β ('softpower').

$$w_{ij} = |corr(X_i, X_j)|^{\beta}$$

with $\beta \geq 1$.

Because softpower is a nonlinear but monotonic transformation of the correlation coefficient, the prediction accuracy measured by AUC will be no different from that of the underlying correlation method itself.

Unsupervised methods: Z-score

Z-SCORE is a network inference strategy by Prill *et al.*
that assumes the availability of **knockout experiments** that
lead to a change in other genes.

The assumption is that the knocked-out gene *i* in experiment *k*
affects more strongly the genes that it regulates than the others.

The effect of gene *i* on gene *j*
is captured with the Z-score z_{ij}:

$$z_{ij} = \left| \frac{x_{jk} - \mu_{X_j}}{\sigma_{X_j}} \right|$$

assuming that the *k*-th experiment is a knockout of gene *i*,
μ_{Xj} and σ_{Xj} are respectively the mean and standard deviation
of the empirical distribution of the expression values x_{jk} of gene *j*.

Unsupervised methods based on mutual information

Relevance networks (RN) introduced by Butte and Kohane measure the **mutual information (MI)** between gene expression profiles to infer interactions.

The MI between discrete variables (here: genes) X_i and X_j is defined as

$$M_{ij} = \sum_{X_i} \sum_{X_j} p(X_i, X_j) \log_2 \frac{p(X_i, X_j)}{p(X_i)p(X_j)}$$

where $p(X_i, X_j)$ is the **joint probability distribution** of X_i and X_j (both variables fall into given ranges) and $p(X_i)$ and $p(X_j)$ are the **marginal probabilities** of the two variables (ignoring the value of the other one).

RELNET

The RELNET is the simplest method based on mutual information.

For each pair of genes, the mutual information M_{ij} is estimated and the edge between genes *i* and *j* is created if the mutual information is above a threshold.

Despite that mutual information is more general than the Pearson correlation coefficient, in practice thresholding the M_{ij} or Pearson correlation produces similar results.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

CLR

The Context Likelihood or Relatedness network (CLR) method is an extension of the RELNET method.

The method derives a score that is associated to the empirical distribution of the mutual information values.

In practice, the score between gene *i* and gene *j* is defined as follows:

$$c_{ij} = \sqrt{c_i^2 + c_j^2}, \text{ with } c_i = \max\left(0, \frac{M_{ij} - \mu_{M_i}}{\sigma_{M_i}}\right) \text{ and}$$

$$c_j = \max\left(0, \frac{M_{ji} - \mu_{M_j}}{\sigma_{M_j}}\right).$$

with the mean μ_{Mi} and standard deviation σ_{Mi} of the empirical distribution of the mutual information between these genes and other genes,

$$\mu_{M_i} = \frac{1}{G}\sum_{l=1}^{G} M_{il}, \quad \sigma_{M_i} = \sqrt{\frac{1}{G-1}\sum_{l=1}^{G}(M_{il} - \mu_{M_i})^2}$$

ARACNE

The motivation of the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) is that many similar measures between variables may be the result of indirect effects.

In order to avoid such indirect effects, the algorithm relies on the "Data Processing Inequality" (DPI).

This approach removes the weakest edge, that is the one with the lowest mutual information, in every triplet of genes.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

PCIT

The Partial Correlation coefficient with Information Theory (PCIT) algorithm combines the concept of partial correlation coefficient with information theory to identify significant gene-to-gene associations.

Similarly to ARACNE, PCIT extracts all possible interaction triangles and applies DPI to filter indirect connections, but instead of mutual information it uses first-order partial correlation as interaction weights.

The partial correlation tries to eliminate the effect of a third gene *l* on the correlation of genes *i* and *j*.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

C3NET

The Conservative Causal Core NETwork (C3NET) consists of two main steps.

In the first step pairwise mutual information is computed.
Then, non-significant connections are eliminated, according to a chosen significance level α, between gene pairs.

In the second step, the most significant edge for each gene is selected.
This edge corresponds to the highest mutual information value among the neighboring connections for each gene.

→ the highest possible number of connections that can be reconstructed by C3NET is equal to the number of genes under consideration.

C3NET does not aim at reconstructing the entire network underlying gene regulation but mainly tries to recover the core structure.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

Feature selection approaches

A GRN reconstruction problem can also be seen as a feature selection problem.
For every gene, the goal is to discover its true regulators among all other genes or candidate regulators. This approach can integrate knowledge about genes that are not TFs and therefore reduce the search space.

Typically, this approach only focuses on designing a significance score *s(i, j)* that leads to a good ranking of the candidate regulations, such that true regulations tend to be at the top of the list since an edge is assigned between *i* and *j* if the evidence *s(i, j)* is larger than a threshold.

With the feature selection approach, the scores *s(i, j)* for all the genes are jointly estimated with a method that is able to capture the fact that a large score for a link *(i, j)* is not needed if the apparent relationship between *i* and *j* is already explained by another and more likely regulation.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

MRNET

The Minimum Redundancy NETworks (MRNET) method reconstructs a network using the feature selection technique known as Minimum Redundancy Maximum Relevance (MRMR), which is based on a mutual information measure.

In order to generate a network, the algorithm performs a feature selection for each gene ($i \in [1, G]$) on the set of remaining genes ($j \in [1, G] \setminus i$).

The MRMR procedure returns a ranked list of features that maximize the mutual information with the target gene (maximum relevance) and, at the same time, such that the selected genes are mutually dissimilar (minimum redundancy).

Bellot *et al. BMC Bioinformatics* (2015) 16:312

MRNET

For every gene, the MRMR feature selection provides a score of potential connections where the higher scores should correspond to direct interactions.

The indirect interactions should have lower scores because they are redundant with the direct ones.

Then, a threshold is computed as in the RELNET method.

The MRNET reconstructs a network using a forward selection strategy, which leads to subset selection that is strongly conditioned by the first selected variables.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

Genie3

The GEne Network Inference with Ensemble of trees (Genie3) algorithm uses the random forests feature selection technique to solve a regression problem for each of the genes in the network.

In each of the regression problems, the expression pattern of the target gene should be predicted from the expression patterns of all TFs.

The importance of each TF in the prediction of the target gene is taken as an indication of an apparent regulatory edge.

Then these candidate regulatory connections are aggregated over all genes to generate a ranking for the whole network.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

supervised inference method: SVM

In contrast to unsupervised methods, e.g. correlation methods, the supervised approach does not directly operate on pairs of expression profiles but on feature vectors that can be constructed in various ways.

E.g. one may use the outer product of two gene expression profiles X_i and X_j to construct feature vectors:

$$\mathbf{x} = X_i X_j^T$$

A sample set for the training of the SVM is then composed of feature vectors $\mathbf{x}_i$
that are labeled $\gamma_i = +1$ for gene pairs that interact and $\gamma_i = -1$ for those that do not interact.

Measure accuracy of GRNs

Inference methods (to infer = *dt. aus etwas ableiten/folgern*) aim to recreate the topology of a genetic regulatory network e.g. based on expression data only.

The **accuracy** of a method is assessed by the extent to which the network it infers is similar to the true regulatory network.

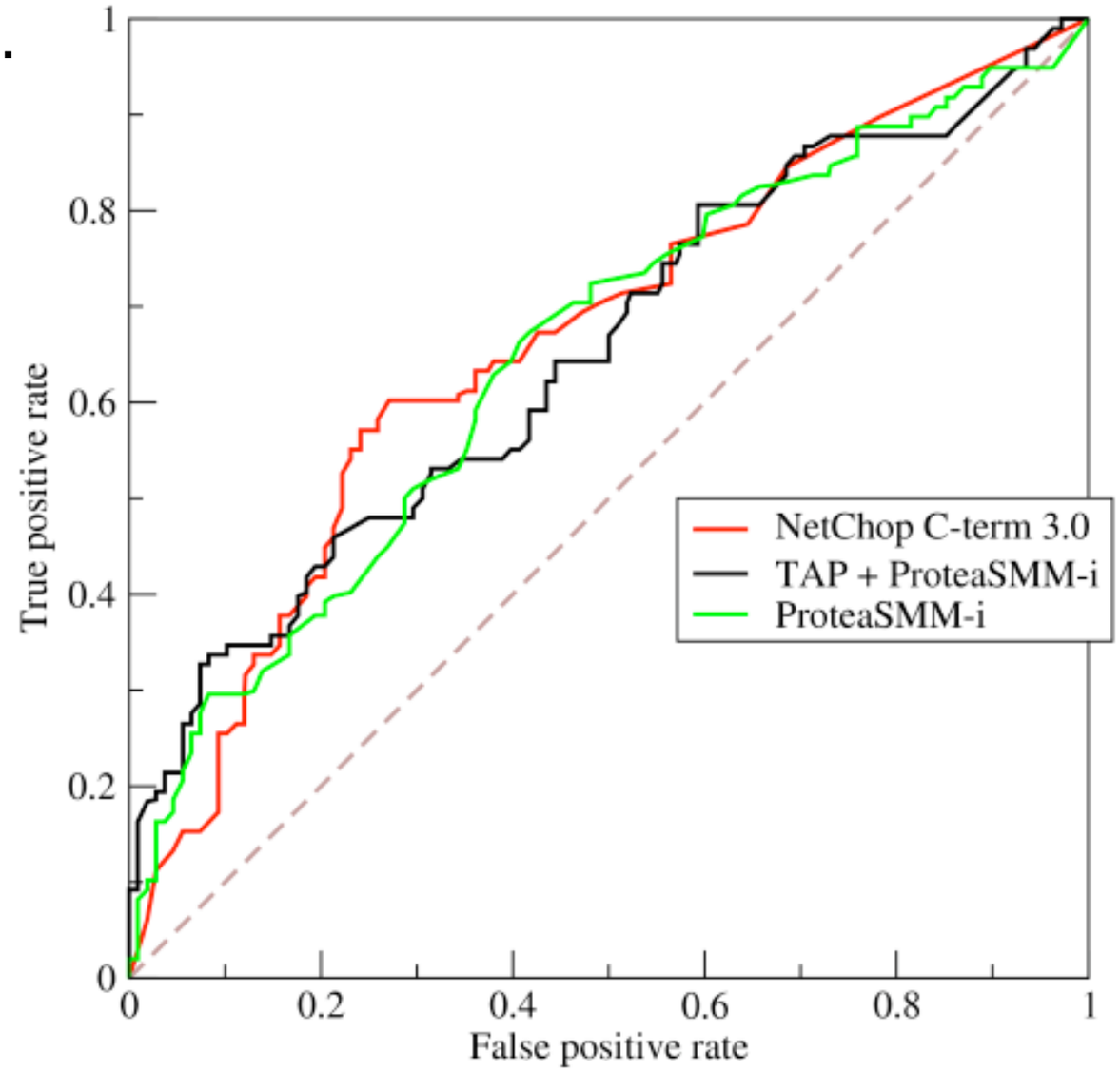
We quantify similarity e.g. by the area under the Receiver Operator Characteristic curve (AUC)

$$AUC = \frac{1}{2}\sum_{k=1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1})$$

where X_k is the false-positive rate and Y_k is the true positive rate for the k-th output in the ranked list of predicted edge weights.

An AUC of 1.0 indicates a perfect prediction, while an AUC of 0.5 indicates a performance no better than random predictions.

AUC



Divide data into bins.

Measure value of function Y at midpoint of bin -> factor 0.5

$$AUC = \frac{1}{2} \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})$$

www.wikipedia.org

Summary

Network inference is a very important active research field.

Inference methods allow to construct the topologies of gene-regulatory networks solely from expression data (unsupervised methods).

Supervised methods show far better performance.

Performance on real data is lower than on synthetic data
because regulation in cells is not only due to interaction
of TFs with genes,
but also depends on epigenetic effects (DNA methylation,
chromatin structure/histone modifications, and miRNAs).

Summary

Today:

- mRNA and protein half-lifes and synthesis rates can be measured experimentally with SILAC MS

- Gene regulation networks have **hierarchies**:
  → global "cell states" with specific expression levels

- Network **motifs**: FFLs, SIMs, DORs are overrepresented
  → different functions, different temporal behavior

- there exist many related methods to generate the topologies of GRNs from gene expression data (warning: weak correlation between TFs and target genes (see V7)

Next lecture:

- benchmarking of GRN methods based on synthetic data