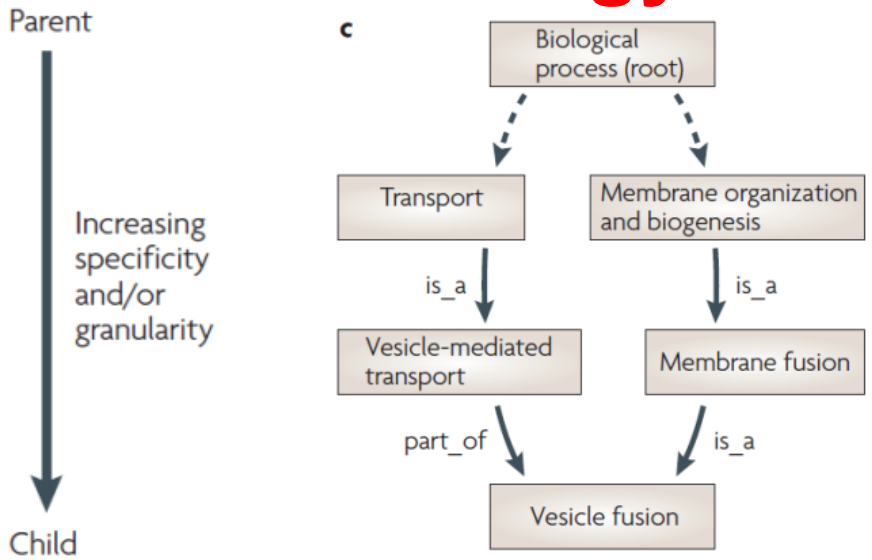
Gene Ontology is a directed acyclic graph



An example of the node
vesicle fusion
in the BP ontology with multiple parentage.

(Arrows point into the wrong direction.)

Dashed edges : there are other nodes not shown between the nodes and the root node.

Root : node with no incoming edges, and at least one leaf.

Leaf node : a terminal node with no children (vesicle fusion).

Similar to a simple tree, a DAG has directed edges and does not have cycles.

Depth of a node : length of the longest path from the root to that node.

Height of a node: length of the longest path from that node to a leaf.

Rhee et al. (2008) Nature Rev. Genet. 9: *509*

relationships in GO

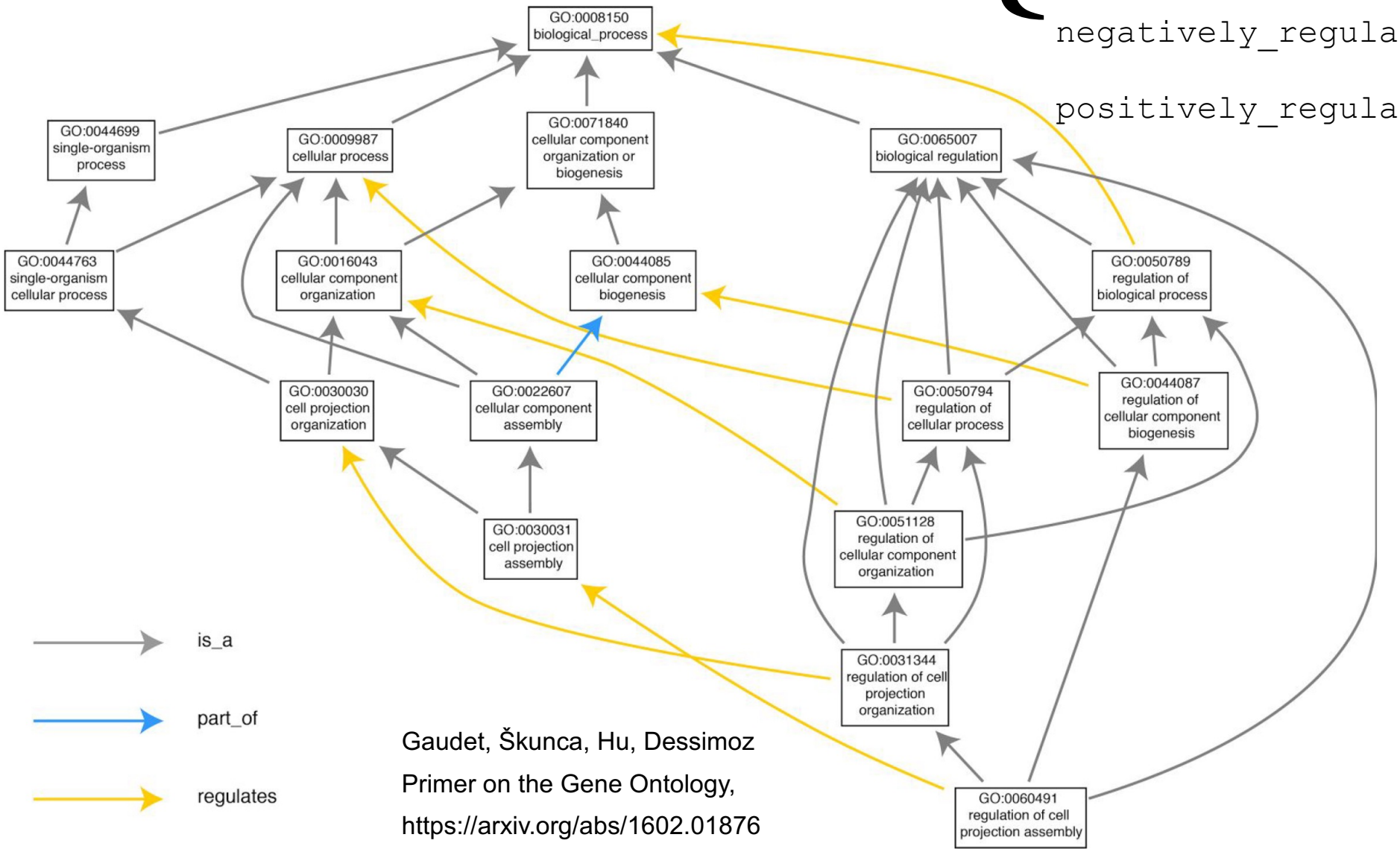
is_a

Gene X { is a part_of

regulates relationship

negatively_regulates

positively_regulates



Gaudet, Škunca, Hu, Dessimoz

Primer on the Gene Ontology,

https://arxiv.org/abs/1602.01876

Hypergeometric test

$$\text{p-value} = \sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

The hypergeometric test is a statistical test.

It can be used to check e.g. whether a biological annotation π is **statistically significant enriched** in a given test set of genes compared to the full genome.

- N : number of genes in the genome
- n : number of genes in the test set
- K_π : number of genes in the genome with annotation π.
- k_π : number of genes in test set with annotation π.

The hypergeometric test provides the **likelihood** that k_π or more genes that were **randomly selected** from the genome also have annotation π.

Hypergeometric test

Select $i \geq k_\pi$ genes with annotation π from the genome.

There are K_π such genes.

The other $n - i$ genes in the test set do NOT have annotation π. There are $N - K_\pi$ such genes in the genome.

$$\text{p-value} = \sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

The sum runs from k_π elements to the maximal possible number of elements.

This is either the number of genes with annotation π in the genome (K_π) or the number of genes in the test set *(n).*

number of possibilities for selecting *n* elements from a set of *N* elements.

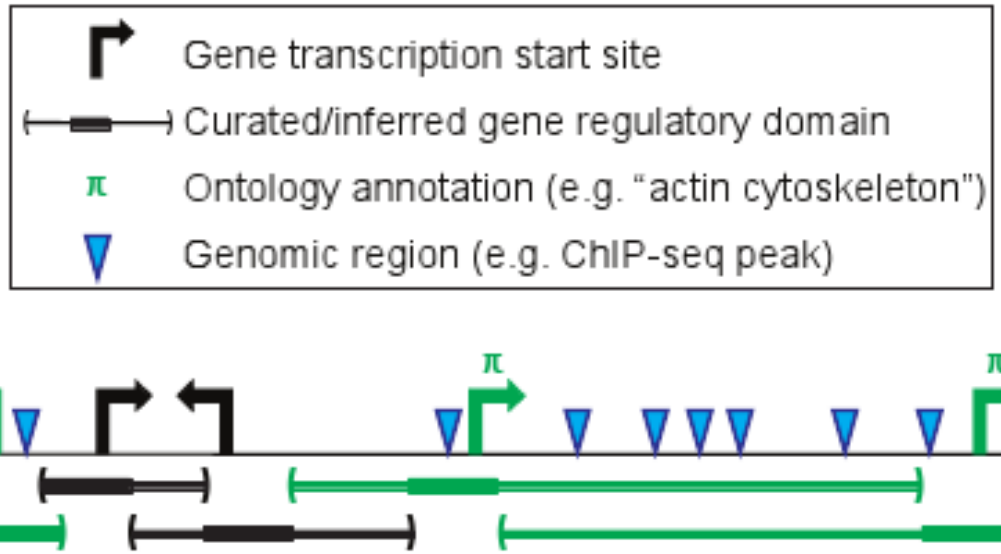
This correction is applied if the sequence of drawing the elements is not important.

http://great.stanford.edu/

http://www.schule-bw.de/

Example

p-value = $\sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$



Is annotation π significantly enriched in the test set of 3 genes?

Hypergeometric test over genes
N = 6 total genes
K_π = 3 genes annotated with π
n = 3 genes with an associated genomic region
k_π = 3 genes annotated and with a genomic region
P-value = 0.05

Yes! p = 0.05 is (just) significant.

http://great.stanford.edu/

Comparing GO terms

The hierarchical structure of the GO allows to compare proteins annotated to different terms in the ontology, as long as the terms have relationships to each other.

Terms located close together in the ontology graph (i.e., with a few intermediate terms between them) tend to be **semantically more similar** than those further apart.

One could simply count the **number of edges** between 2 nodes as a measure of their similarity.

However, this is problematic because not all regions of the GO have the same **term resolution**.

Gaudet, Škunca, Hu, Dessimoz
Primer on the Gene Ontology,
https://arxiv.org/abs/1602.01876

Information content of GO terms

The **likelihood** of a node *t* can be defined in 2 ways:

How many genes have annotation *t* relative to the root node?

$$p_{anno}(t) = \frac{\cdot\, occur(t)}{occur(root)}$$

Number of GO terms in subtree below *t* relative to number of GO terms in tree

$$p_{graph}(t) = \frac{D(t)}{D(root)}$$

The likelihood takes values between 0 and 1 and increases monotonic from the leaf nodes to the root.

Define **information content** of a node from its likelihood:

$$IC(t) = -\log p(t)$$

A rare node has high information content.

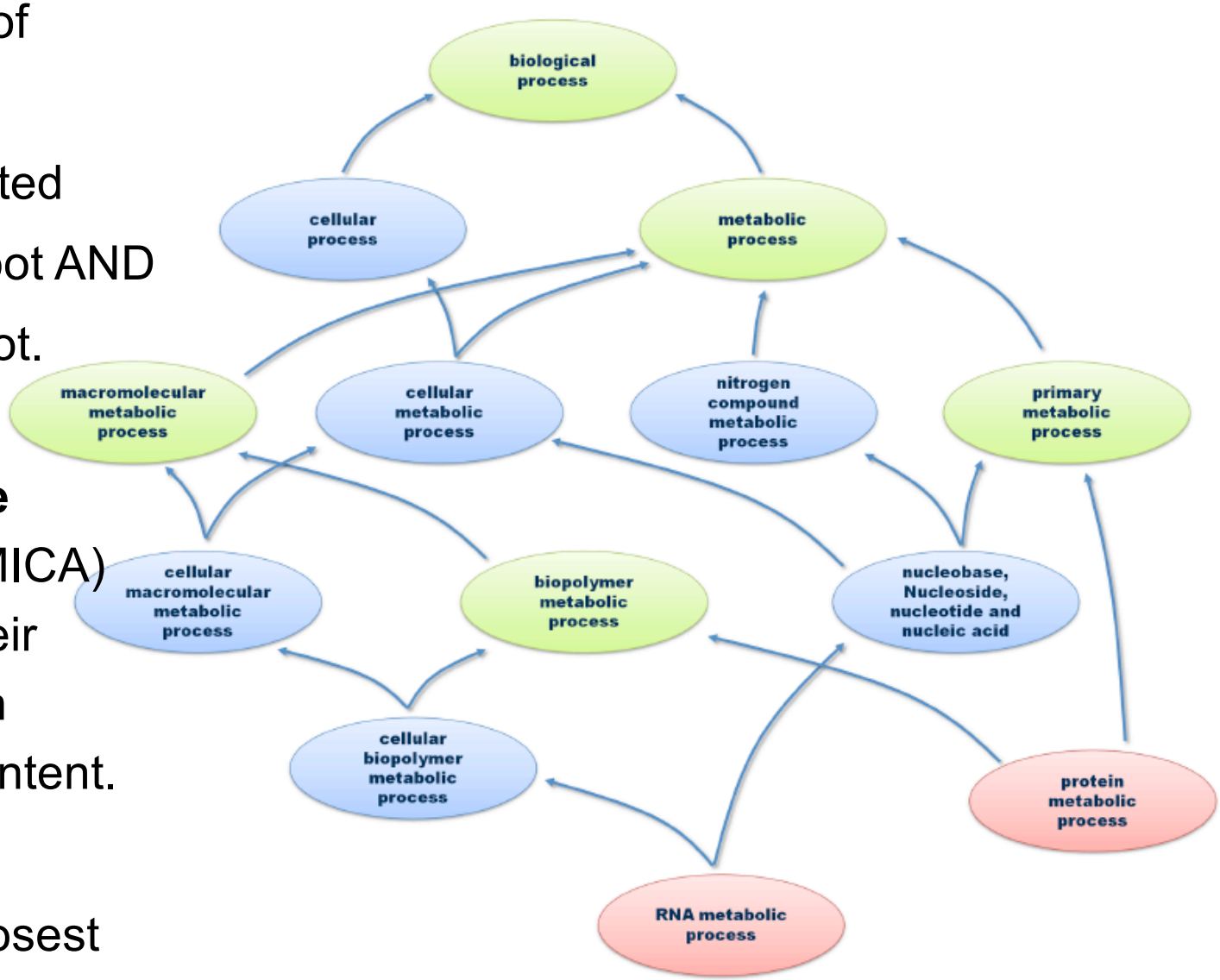
PhD Dissertation Andreas Schlicker (UdS, 2010)

Common ancestors of GO terms

Common ancestors of two nodes t_1 and t_2 : all nodes that are located on a path from t_1 to root AND on a path from t_2 to root.

The **most informative common ancestor** (MICA) of terms t_1 und t_2 is their common ancestor with highest information content.

Typically, this is the closest common ancestor.



PhD Dissertation

Andreas Schlicker (UdS, 2010)

Measure functional similarity of GO terms

Lin *et al.* defined the **similarity** of two GO terms t_1 und t_2

based on the information content of the most informative common ancestor (MICA)

$$sim_{Rel}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1)+IC(t_2)}$$

If MICAs are close to the two GO terms, they receive a high similarity score.

Schlicker *et al*. defined the following variant:

$$sim_{Rel}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1)+IC(t_2)} \cdot (1 - p(MICA))$$

where the term similarity is weighted with the counter-probability of the MICA.

By this, shallow annotations (low “depth” in the tree, slide #4) receive less relevance than MICAs further away from the root.

PhD Dissertation Andreas Schlicker (UdS, 2010)

Measure functional similarity of two genes

Two genes or two sets of genes *A* und *B* typically have more than 1 GO annotation each. → Consider similarity of all terms *i* and *j*:

$$s_{ij} = sim(GO_i^A, GO_j^B), \forall i \in 1,...,N, \forall j \in 1,...,M.$$

and select the maxima in all rows and columns:

$$rowScore(A,B) = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \le j \le M} s_{ij}, \qquad GOscore_{avg}^{BMA}(A,B) = \frac{1}{2} \cdot (rowScore(A,B) + columnScore(A,B))$$

$$columnScore(A,B) = \frac{1}{M}\sum_{j=1}^{M} \max_{1 \le i \le N} s_{ij}. \quad GOscore_{max}^{BMA}(A,B) = max(rowScore(A,B), columnScore(A,B))$$

Compute *funsim*-Score from scores for BP tree and MF tree:

$$funsim(A,B) = \frac{1}{2} \cdot \left[\left(\frac{BPscore}{\max(BPscore)} \right)^2 + \left(\frac{MFscore}{\max(MFscore)} \right)^2 \right]$$

PhD Dissertation Andreas Schlicker (UdS, 2010)

Rates of mRNA transcription and protein translation

ARTICLE

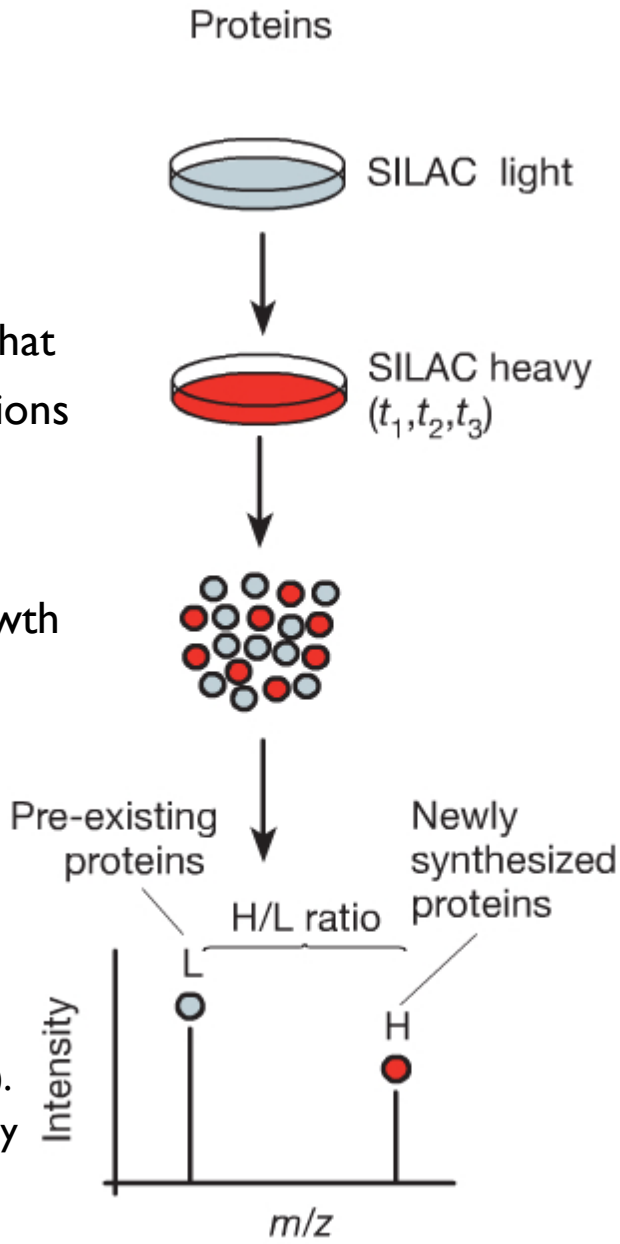
doi:10.1038/nature10098

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.



Quantification of protein turnover and levels. Mouse fibroblasts were pulse-labelled with heavy amino acids (SILAC). Protein turnover is quantified by mass spectrometry.

Schwanhäuser et al.
Nature 473, 337 (2011)

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]

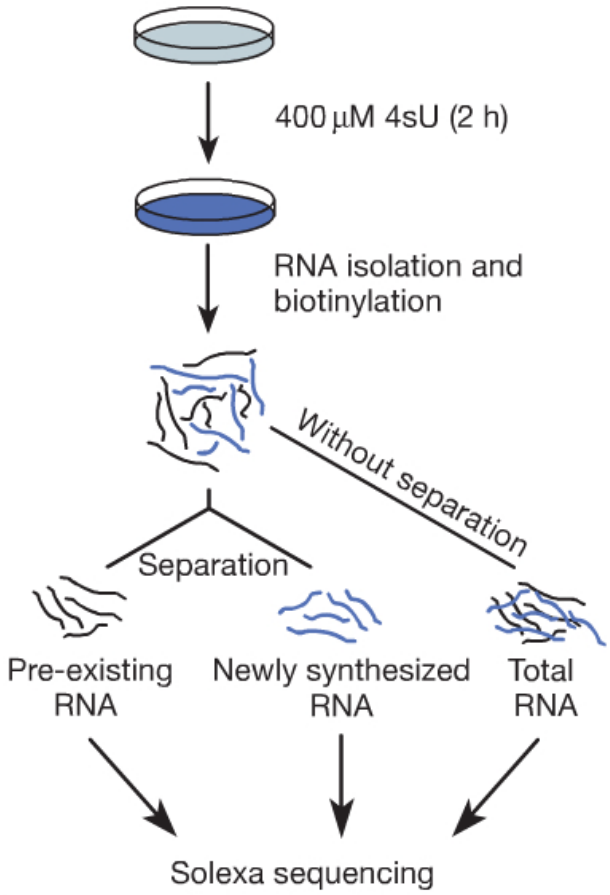
Quantification of mRNA turnover and levels. Mouse fibroblasts were pulse-labelled with the nucleoside **4-thiouridine** (4sU). mRNA turnover is quantified by next-generation sequencing.

The 4sU-labeled RNA fraction is thiol-specifically biotinylated generating a disulfide bond between biotin and the newly transcribed RNA.

'Total cellular RNA' can then be quantitatively separated into labeled ('newly transcribed') and unlabeled ('pre-existing') RNA with high purity using streptavidin-coated magnetic beads.

Finally, labeled RNA is recovered from the beads by simply adding a reducing agent (*e.g.* dithiothreitol) cleaving the disulfide bond and releasing the newly transcribed RNA from the beads.

Rädle, J Vis Exp. 2013; (78): 50195.



Schwanhäuser et al. Nature 473, 337 (2011)

Rates of mRNA transcription and protein translation

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

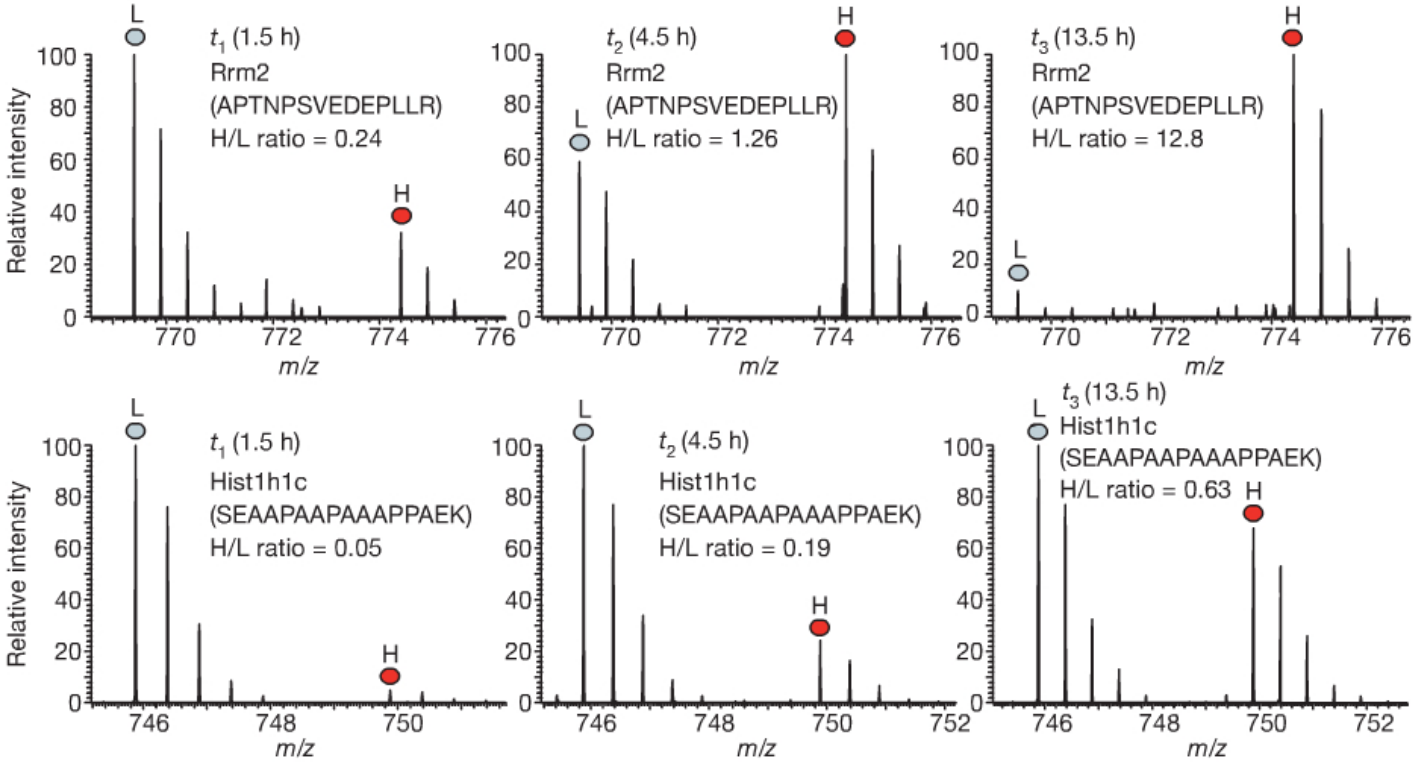
5,279 of these proteins were quantified by at least 3 heavy to light (H/L) peptide ratios belonging to these proteins.

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

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

You should understand these spectra!

Top: **high-turnover protein**



Schwanhäuser et al. Nature 473, 337 (2011)

Bottom: **low-turnover protein**, slow synthesis, long half-life

Protein half-lifes and decay rates

Consider 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_∞):

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

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

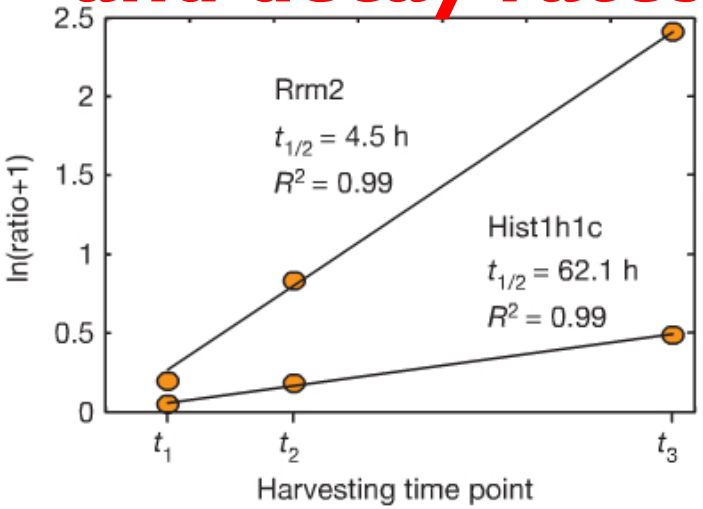
$$\frac{P_H}{P_L} + 1 = \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}}$$

take *ln* on both sides

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

$$ln(ratio + 1) = k_{dp} t + \frac{t}{t_{cc}} ln2 = t \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right) ln(ratio + 1)\, t = t^2 \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right)$$

$$ln(ratio + 1)\, t = t^2 \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right)$$



Consider m intermediate time points:

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

From k_{dp} we get the desired half-life:

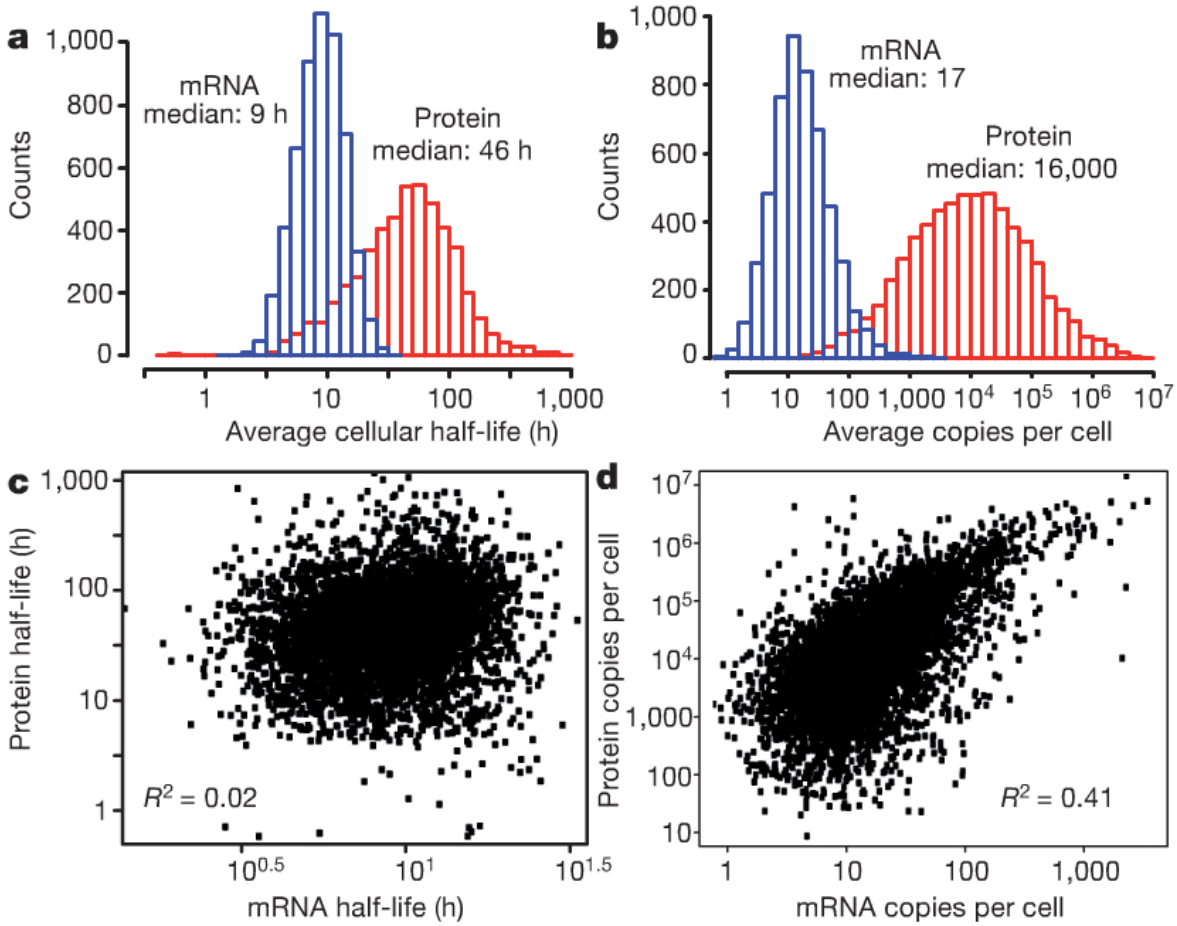
$$T_{1/2} = \frac{\log_e 2}{k_{dp}} .$$ because this gives

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

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).

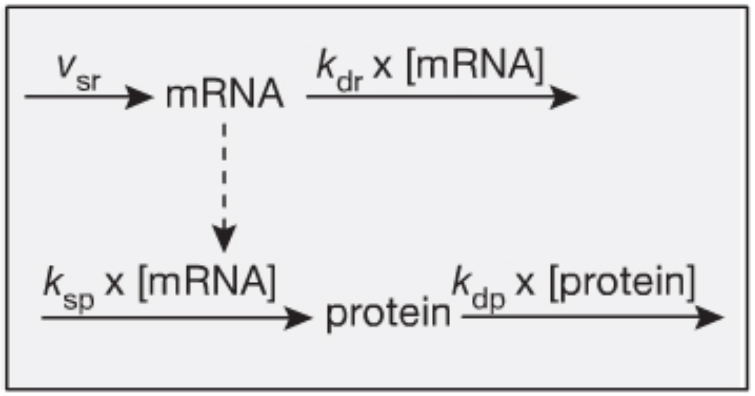
Proteins were on average 5 times more stable (46h vs. 9h) and 900 times more abundant than mRNAs.

(right) mRNA and protein levels showed reasonable correlation ($R^2 = 0.41$)
(left) However, there was practically no correlation of protein and mRNA half-lives.

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

Top

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

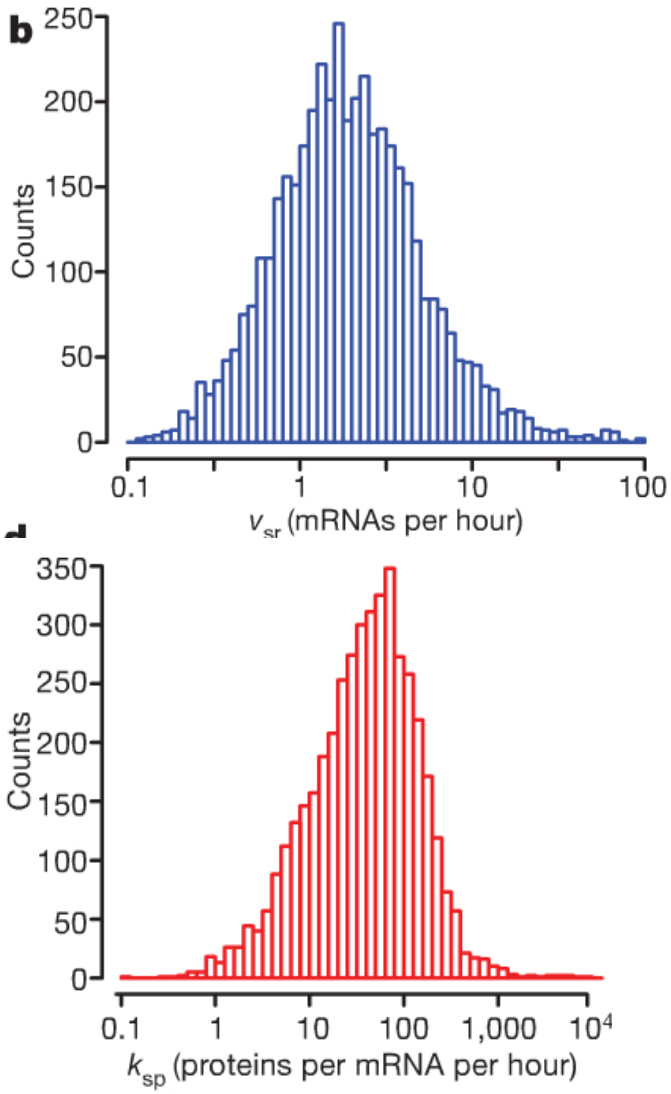
The median is about 2 mRNA molecules per hour (**very slow**!).

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

Bottom

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

Schwanhäuser et al. Nature 473, 337 (2011)



Calculated translation rate constants are not uniform