# How Little Do We Actually Know? – On the Size of Gene Regulatory Networks.

Richard Röttger, Jan Baumbach
*Max Planck Institute for Informatics, Saarbrücken, Germany*

roettger@mpi-inf.mpg.de

**Abstract:** Nowadays, we have whole-genome sequences for more than more than two thousand species available for download from the NCBI databases. Ongoing improvement of DNA sequencing technology will further feed this trend. However, the availability of sequence information is only the first step in understanding how cells survive, reproduce and adjust their behavior. The molecular biological mechanisms, which control organized development and adaptation of complex organisms still remain widely undetermined. Transcriptional gene regulation is one of the key players here. The direct juxtaposition of the total number of sequenced species to the handful of model organisms with known regulations is astonishing. Recently, we investigated how little we even know about these model organisms. Our aim was to predict the sizes of the whole-organism regulatory networks of seven species. In particular, we provided a statistical lower bound for the expected number of regulations. For *E. coli* we estimate at most 37% of the expected gene regulatory interactions to be already discovered, 24% for *B. subtilis*, and < 3% for human respectively. We conclude that even for our best-researched model organisms we still lack substantial understanding of fundamental molecular control mechanisms, at least on a large scale.

Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/22585140
Web: http://dx.doi.org/10.1109/TCBB.2012.71

## 1 Introduction

In 2010, whole-genome sequences of more than 1200 microbes, 3600 viruses and 39 eukaryotic species, as well as over 2400 sequences for eukaryotic organelles were available for download at the web site of the National Center for Biotechnology Information, NCBI [ea11]. The continuously improving next-generation sequencing techniques will further increase the number of sequenced species dramatically. The heavily reduced cost for DNA sequencing by over two orders of magnitude allows us to utilize

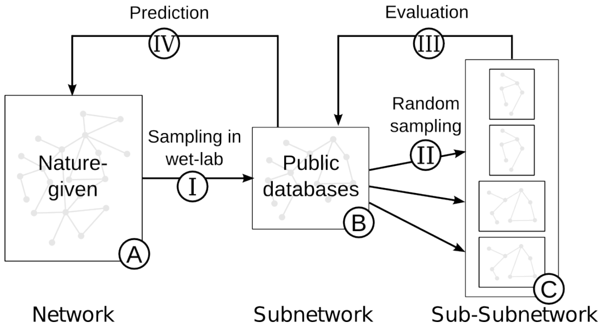

Figure 1: Overview: We aim to infer the size of the nature-given transcriptional gene regulatory network (A) of a specific organism. Given is a known subnetwork (B), which was sampled by biologists in laboratories and stored in public databases (I). In order to study the reliability of our prediction, we randomly sample (II) even smaller networks, subsubnetworks (C) and subsequently test if we can predict the known network size and evaluate the robustness of the approach (III). After verifying the estimator to be bias-free and robust, we are able to predict (IV) the size of the nature-given network (A).

this technology as routine method for unraveling the genetic repertoire of numerous species of varying complexity and ecological, economic and medical significance [Met10].

However, the availability of genome sequences, even complete ones, is only the first step towards a comprehensive understanding of how cells survive, reproduce and adapt to changing environmental conditions. One major molecular control mechanism of cells is transcriptional gene regulation. Key players are the so-called transcription factors (TFs), proteins that possess DNA-docking domains to bind certain regions within the DNA sequence. Thereby they influence the expression of numerous target genes (TGs) and control genetic programs like growth, survival, reproduction, digestion, immune responses, etc. Transcriptional gene regulatory networks (GRN) emerge, with nodes corresponding to genes and directed edges that represent regulatory interactions [PS92].

Understanding this fundamental molecular control mechanism on a large scale is one of the most important goals in systems biology and a prerequisite for the subsequent modeling and analysis of cell response and behavior [BWKT09]. However, our current knowledge is limited and the

| **Organism** | $\mathbf{E^S}$ | $\widehat{\mathbf{E}}^{\mathbf{N}}$ | **95% Lo.** | **Ratio** | **Ratio CI** |
|---|---|---|---|---|---|
| *H. sapiens* | 3,902 | 165,807 | 130,326 | 2.35% | 2.99% |
| *M. musculus* | 1,730 | 190,359 | 157,251 | 0.91% | 1.08% |
| *R. norvegicus* | 804 | 421,819 | 313,890 | 0.19% | 0.26% |
| *A. thaliana** | 1,852 | 569,013 | 111,026 | 0.31% | 1.67% |
| *E. coli* | 3,946 | 15,399 | 10,562 | 25.63% | 37.36% |
| *B. subtilis* | 1,391 | 9,716 | 5,780 | 14.31% | 24.07% |
| *C. glutamicum* | 806 | 9,114 | 5,696 | 8.84% | 14.15% |

*this is a slightly modified dataset. Please refer to the main paper for an exhaustive discussion.

Table 1: The table shows a summary of our results for four species. Most important are the last columns Ratio and Ratio CI, which give the fraction of known regulatory interactions in relation to the predicted nature-given network size $\widehat{E}^N$, i.e. an estimation of how much we have discovered yet. In the column Ratio, we use the estimation $\widehat{E}^N$ for the comparison, whereas in Column Ratio CI the lower bound of the 95% confidence interval was used (95% Lo.). The column $E^S$ gives the number of known regulations for each species.

reconstruction of the gene regulatory networks is far from being complete. Even for *E. coli*, the model organism with the largest currently available experimentally validated data for any free-living organism, we have information about the transcriptional regulation of only around one third of the genes [S. 08].

With this short highlight paper, we briefly describe (1) a robust model to estimate the expected size of transcriptional gene regulatory networks, i.e. the number of edges, and (2) conclude that we actually know very little about one of the most important mechanisms that controls genetic activity.

## 2 Model and Results

Our estimator model essentially works on graph invariants. GRNs are directed graphs with two types of nodes (genes): the transcription factors (TFs) and the target genes (TGs). With our network size prediction model we account for the two types of nodes but also for three types of regulatory interactions (edge types): (TF→TF) regulations between two transcription factors, (TF→TG) regulations between a transcription factor and target gene and (TF-self) self regulations of transcription factors.

Subsequently, for each species with a partially known GRN, we calculated the three edge probabilities. As these probabilities form graph invariants, we can use them to estimate the total size of the GRN. Two challenges arise: (1) We lack a gold standard, i.e. a fully known GRN for at least one organism, that we could use for validating our model. (2) We usually have only one reference database per species that describes the known GRN of this organism. Thus assessing the variability of our method is not straight forward. Figure 1 gives an overview about the utilized methodology. For robustness assessment we use bootstrapping methods to receive confidence intervals for our estimations. Finally, we were interested in the lower bounds, i.e. how much of an organism's GRN can we expect to know in the very best case. We applied our model to seven species, among them human, mouse, *E. coli* and *B. subtilis*. See Table 1 for our major results for these four species. We found it quite astonishing that even for *E. coli* only 37% of the expected GRN is known, in the best case. For mammals our current knowledge is too limited to even give trustworthy estimations.

For formal definitions and exhaustive description of the mentioned methods, please refer to the main paper's method section [RRTB12].

## References

[BWKT09] J. Baumbach, T. Wittkop, C.K. Kleindt, and A. Tauch. Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nature Protocols*, 4(6):992–1005, 2009.

[ea11] E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011.

[Met10] M. L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[PS92] C.O. Pabo and R.T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61(1):1053–1095, 1992.

[RRTB12] R. Rottger, U. Ruckert, J. Taubert, and J. Baumbach. How Little Do We Actually Know? – On the Size of Gene Regulatory Networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, PP(99):1, 2012.

[S. 08] S. Gama-Castro et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*, 36(Database issue):D120–4, 2008.