Project Details

An attachment for the GAČR project proposal
Uncovering Regulatory Interactions in Cells with Bayesian Statistics

Biological Background

One of the great feats of modern biology was the elucidation of the central dogma of molecular biology describing the relationship between the information-carrying molecules at the base of all life: DNA, RNA and proteins. In a very simplified form the central dogma may be paraphrased as follows: The DNA contains all of the instructions needed to run a cell and can be divided into *coding* and *non-coding* regions. The coding regions of DNA can be *transcribed* to form *messenger RNA* (mRNA) molecule containing a mirror-copy of the coding region. The messenger RNA is then *translated* to create protein. Proteins then perform most of the actual functions of the cells, including control of transcription and translation. This process is regulated at all stages and involves many feedback loops: most importantly, the rate of transcription of a gene is controlled by the abundance of regulatory proteins binding to specific sequences of the DNA near the coding region. The rate of translation and degradation of mRNA can also be regulated separately by other proteins and the rate of degradation of proteins themselves is also affected by other proteins. There are many exceptions where the above simplification does not hold, but the vast majority of cellular life can be described in these terms.

The complex web of regulatory interactions and feedback loops forms the genetic program of the cell that directly or indirectly determines cellular behavior. Elucidating those interactions thus means understanding how the genetic code is “executed” with implications for medicine, biology and biotechnology. This is however not a simple task: even in the simplest bacteria there are hundreds of genes that give raise to thousands of proteins and their variants and thousands of interactions, while the human genome contains tens of thousands of genes that give raise many more proteins and their interactions.

One of the challenges in understanding the genetic program is that we are generally unable to predict any of the aforementioned processes and interactions from the DNA sequence alone. Moreover, even our ability to observe what happens in the cell is very limited: we are unable to directly measure concentrations of more than several hundred proteins at a time, and determining the segments of the DNA where proteins bind has to be performed for each protein separately. The only commonly available method that can capture information about all genes in a single experiment is measuring the concentration of mRNA which is in turn strongly related to *expression* (how many mRNAs are transcribed from the gene in a unit of time). Per-gene concentration of mRNA can be measured either through microarrays or RNA-seq. The former is an older and less flexible technology and the latter is the state of the art. Since large amounts of microarray data are available, microarrays remain an important source of information. Both methods to gather expression data are relatively expensive and except for the most studied organisms, at most several dozen whole genome expression experiments have been published.

It has been shown that mRNA and protein concentration are mostly correlated, however this relationship is imperfect (Maier et al. 2009; Gygi et al. 1999). Nevertheless, expression

data is the best proxy for protein concentration available at the whole genome level and expression data are thus widely used to infer interactions that control transcription of genes.

Current State of Knowledge

In inferring gene regulatory interactions there are two large groups of approaches: *model-based* and *model-free*. The former try to use an explicit, biologically plausible model of the regulatory interactions or an approximation thereof, while the latter rely on more general techniques from statistics or machine learning to find associations without directly taking into account the underlying biological processes. While there is little doubt that the model-based approach is ultimately more correct, the available data are usually insufficient to determine parameters of high fidelity models and some additional assumptions about the underlying biological processes have to be made to form feasible approximations. To this end model-free approaches are sometimes chosen because they can process larger amounts of data while making fewer assumptions about the underlying biology. In this project we focus on improving a model-based approach, but we will discuss state of the art in model-free methods as well. The literature on the topic is vast, in the sections below we report only selected methods that represent the individual directions that are present in the literature.

Model-Free Methods

Most model-free methods are derived from machine learning algorithms. One of the most successful is GENIE 3 which is based on random forests (Huynh-Thu et al. 2010). GENIE 3 aims to predict the expression level of each target gene separately based on expression of candidate regulators. The candidate regulators are then ordered by their average influence in all of the decision trees built. GENIE 3 assumes the data come from the same experiment or that they have been properly normalized, but makes no additional assumptions.

The NetworkBMA algorithm (Yeung et al. 2011) uses linear regression with time series data, using expression of other genes at a previous time point as predictors and treats interaction inference as variable selection on top of this linear model.

More modern machine learning methods have been tried, including deep learning (Chen et al. 2016) or boosting with regression stumps (Mall et al. 2018) but those have not become a standard tool in analysis.

Another popular class of approaches is rooted in information theory. In particular high mutual information between the putative regulator and the target is used as an indicator that the regulation is taking place. This is represented by the ARACNE family of algorithms (Margolin et al. 2006; Zoppoli et al. 2010; Lachmann et al. 2016). The ARACNE family makes similar assumptions as GENIE 3.

The main issue with all model-free methods is that they often have difficulties inferring the direction of regulatory relationships and that it is usually difficult to interpret their results biologically.

Model-Based Methods

The model-based approaches span a spectrum from binary models where the genes are only considered active or inactive (e.g. Akutsu et al. 1999) to simulating full Michaelis-Menten kinetics for all reactions in the transcription process (Goutsias and Lee 2007). While binary models are now mostly considered too simplistic, using full kinetics remains impractical. The

currently most used models stay in the middle of this spectrum: they assume that the concentration of mRNA is determined either by a linear ordinary differential equation (ODE) or by ODE with a sigmoid non-linearity on top of a linear combination of the regulators (Vohradský 2001). It can be shown that under realistic assumptions, the sigmoid model is a good approximation to modelling the transcription process completely (Veitia 2003). The sigmoid models are currently the most complex that can be reliably identified from the data available.

In this context the ODE parameters can be fit in two ways: when time-course data are available, the ODE can be solved numerically and parameters optimized directly. A simpler option is to perform gradient matching - use the differences between successive time points as approximate derivatives and optimize the ODE parameters to fit those derivatives. When the available data are not time-course, gradient matching is the only option and it is usually assumed that the organism is in *steady state* meaning that all the derivatives are approximately zero.

The Inferelator algorithm (Bonneau et al. 2006) uses gradient matching and a linear model to combine steady-state and time-series data from a large number of experiments. EGRIN-2 (Brooks et al. 2014) is a huge database of possible regulatory interactions in *Escherichia coli* and *Halobacterium salinarum* built from a large and diverse set of expression experiments derived using Inferelator and cMonkey biclustering algorithm (Reiss et al. 2006). In this workflow, cMonkey groups together both genes and experiments which have similar expression patterns and the inference is then performed on the resulting clusters to reduce dimensionality.

The H-Licorn algorithm (Chebil et al. 2014) uses an approximate three-state discrete model as a preprocessing step to generate candidate regulations and then fits a linear model via gradient matching to the candidate regulations and keeps the one with lowest prediction error. This two-step process is then performed repeatedly in a bagging scheme.

(Berrones et al. 2016) propose a method tailored for periodic regulations (e.g., the circadian clock) where the expression time series is decomposed via Fourier transformation and gradient matching is used to fit model parameters.

A recurring theme in more recent work is using regulations known from literature and fitting the model for known regulations first, treating the expression of the regulators as random variables. This has the effect of denoising the expression of the candidate regulators and their inferred expression values are then used to predict the other target genes (Arrieta-Ortiz et al. 2015; Zhang et al. 2013; Fu et al. 2011).

Bayesian Model-Based Methods

The basis of a fully Bayesian approach to model fitting is to quantify uncertainties in model parameters with posterior probability distributions instead of just finding point or interval estimates. In most cases, Bayesian approach entails that a *generative model* is used - such a model is a full probabilistic description of the process that is assumed to have generated the observed data from the unknown parameters. When a generative model is given, numerical methods based on Markov chain Monte Carlo are able to find posterior distributions automatically or semi-automatically. The basic models used in the Bayesian settings are the same as in the previous section, except they are augmented with priors on parameters to make the models generative.

An early example is (Morrissey et al. 2010) where the interaction network is modelled as a linear autoregressive process that is fit with a Bayesian approach. Pioneering work by

(Titsias et al. 2012) uses the sigmoid model of regulation combined with time series data and explicit solving of the ODEs in a fully Bayesian setting. This approach has theoretical appeal, but a major downside are its huge computational requirements. More recently BGRMI (Iglesias-Martinez et al. 2016) uses a linear model with time series and gradient matching in a Bayesian setting.

Our Contribution

We have worked on modelling transcriptional regulation with focus on *Bacillus subtilis*, including biological validation of the results (Ramaniuk et al. 2017). We have also implemented open source software that uses maximum-likelihood estimation for the sigmoid model with explicit ODE solving (Modrák and Vohradský 2018).

Aims & Scope

This project aims to take one step forward in our ability to infer transcriptional regulations and help the scientific community better understand how cells function. Since the available data mostly have few observations per gene, taking measurement uncertainty explicitly into account holds promise for improved inference. This is exactly the appeal of Bayesian methods which handle the uncertainty to a full extent. One of the main arguments against Bayesian methods is that they are feasible only for small datasets. However recent advances in Monte Carlo methods for Bayesian inference (Carpenter et al. 2017) allowed speedups that should make Bayes feasible for many practical datasets.

Another advantage is that formulating a generative Bayesian model provides principled way to combine steady-state data with time series while solving ODEs explicitly (e.g., without resorting to gradient matching). This is simply a side effect of having a generative model for both the time-series and steady-state data.

In the same vein, having an explicit generative model makes it possible to combine multiple experiments, even when they were performed with a different method (RNA-seq vs. microarray), we need to formulate a different measurement error model and find a way to express different experiments in terms of single underlying regulatory network. The latter task requires at least to fit a special normalization factor for each experiment, as the total amount of mRNA captured certainly differs.

At present time, we have a prototype that matches the state of the art in Bayesian methods for time-series data (see Figure 1). The model is also able to infer denoised expression profiles of a set of regulators from known regulations and then use this profile to infer novel regulations. The prototype is implemented in the Stan probabilistic programming language and embedded in an R package, the source code of the prototype is available at https://github.com/cas-bioinf/genexpi/tree/master/genexpi-stan

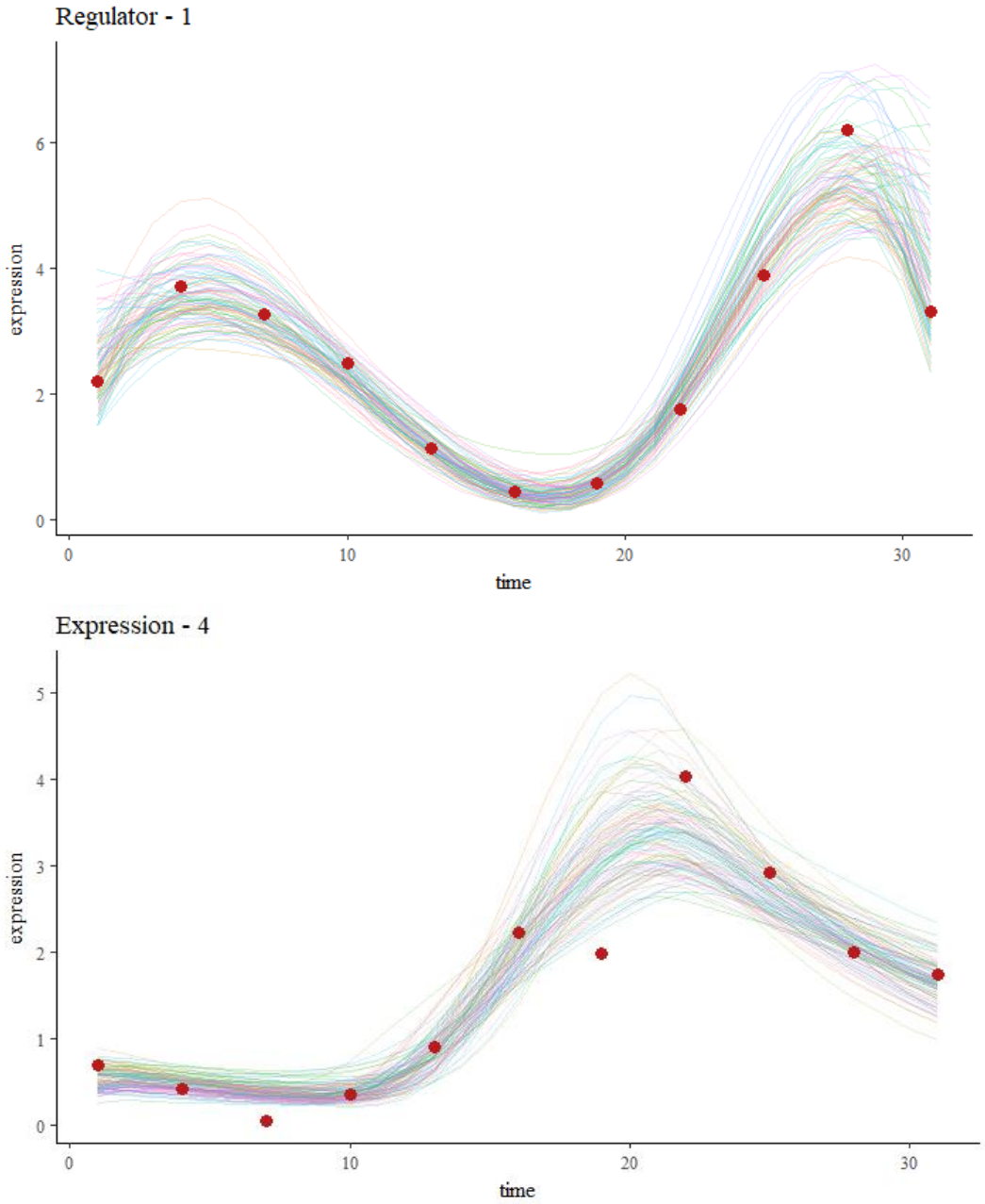

Figure 1: Samples from posterior distribution of expression levels (lines) and the observed expression values (dots) for a regulator (top) and a possible target gene (bottom).

Overall, our aim is to extend and improve this model in four parts:

Infrastructure: Allow the model to combine multiple experiments, especially handling normalization across the experiments

Step 1: Extend the model to handle combined time series and steady state data

Step 2: Allow the model to handle combined microarray and RNA-seq data with separate error models for each

Step 3: Relax the assumption that the regulatory interactions are exactly the same across all experiments.

All of the steps rely on infrastructure. The ordering of Step 1 and 2 is determined by the fact that there are seldom multiple time series data for a single organism and thus it makes sense to first support steady state before supporting multiple experiment types. To our knowledge, there is no Bayesian inference tool that would satisfy any of these steps individually, let alone combined.

In each step, there are multiple connected goals:

- **Develop and implement the improved model.** The model will be implemented in the Stan probabilistic language and wrapped in a package for the R language. All source code will be open and deposited at GitHub or other suitable open-source repository.
- **Gather and prepare data to be used with the model.** While expression data from published works are generally available, they are stored in a variety of formats and usually only processed normalized data and raw data is available. To model the measurement error well, we need processed, but unnormalized data and thus we might need to reprocess the raw data. Due to the large variety of tools and data formats in use, the reprocessing step is non-trivial and may take significant time.
- **Validate the model computationally.** This involves both studies with simulated data, where the true regulations are known with certainty, and studies with real data, comparing the regulatory interactions determined by the model with known regulations described in the literature. For steps 1 & 2 we will also compare the results of our method to other tools that infer regulations from the same types of data. This might not be possible for step 3 as very few methods are developed for this scenario and those may not be easily available to execute on our data. We will also adhere to the principles of reproducible open science - all results will be made available as a part of executable notebooks, letting anyone rerun or modify our analysis.
- **Apply the model.** Once the model is validated it can be used to predict new possible regulations from the available data. Biological interpretation using expert knowledge of the organism in question plays an important part in making the results useful to the broader scientific community. Auxiliary analyses (e.g. consensus sequences of gene promoters, pathway enrichment …) will be run as needed.
- **Validate the model biologically.** The ultimate test of our model will be its biological relevance. It is therefore vital to see if the previously unknown regulatory interactions identified by the model really take place in the studied organism. The primary methods we plan to use are *in-vitro* transcription systems where a DNA template with the target gene promoter, RNA polymerase holoenzyme containing the predicted regulator (sigma factor) are mixed together and the concentration of transcribed mRNA is measured. Due to its cost in both time and material, biological validation will be performed only for a small number of predicted regulations.

The biological validation will be primarily focused on the bacteria *Bacillus subtilis* and *Mycobacterium smegmatis* as we have previous experience with those organisms. Possible other organisms for computational validation will be determined based on availability of suitable expression data.

Schedule

Table 1 provides a schedule and time requirements for the individual steps and related activities. The goals for each step are generally interleaved and cannot be ordered within each year.

Task	2019	2020	2021	Total
Infrastructure implementation	3	-	-	**3**
Step 1 implementation	2	-	-	**2**
Step 2 implementation	-	3	-	**3**
Step 3 implementation	-	-	3	**3**
Data acquisition and preprocessing	2	2	-	**4**
Computational validation	2	2	2	**6**
Applying the model and biological interpretation of the results	1	1	2	**4**
Biological validation of selected results	2	2	2	**6**
Publication and dissemination	-	1	2	**3**
Testing and improving usability of the software	-	1	1	**2**
Total	**12**	**12**	**12**	**36**

Table 1: Schedule of the activities. All time units are in man-months.

Collaboration with Foreign Institutions

We have setup an informal collaboration with the development team of the Stan probabilistic programming language which is located primarily at Columbia University, NY, USA with additional members across the USA and at Aalto University in Finland. (see http://mc-stan.org/about/team/ for a full list).

In this collaboration we work with the Stan team to fix issues and add features for the Stan language and accompanying tools to help us with the development of our models. This collaboration is expected to be maintained throughout the project duration.

Facilities

The project will be executed by Laboratory of Bioinformatics with the help of the Laboratory of Microbial Genetics and Gene Expression at the Institute of Microbiology of the Czech Academy of Sciences.

The Laboratory of Bioinformatics has experience in developing models of transcriptional regulation (e.g., Vu and Vohradsky 2007; Vohradsky 2012) and we have developed a prototype of a Bayesian model of regulation as described in the Aims & Scope section. The

Laboratory of Bioinformatics is part of the Elixir infrastructure which provides access to large computing clusters. These clusters may be necessary for thorough computational validation of the method.

The Laboratory of Microbial Genetics and Gene Expression has experience working with both *Bacillus subtilis* and *Mycobacterium smegmatis* including well proven in vitro transcriptional systems that are a good tool to evaluate the results of the model biologically. The laboratory also has the necessary expertise to interpret the results of the model and guide its development to stay true to the actual biological processes in the cell.

Team

Martin Modrák, PhD. will be responsible for design and implementation of the method, performing evaluation and validation and for managing the project. He is an accomplished computer scientist and programmer with experience in both academia and industry.

We will also partially employ one postdoc from the laboratory of Microbial Genetics and Gene Expression, responsible for consulting on the biological relevance of the model, biological interpretation and validation of the results.

We will further employ two diploma students that would help with data gathering, applying the model on new data, auxiliary analyses, validation of the model, and software testing and quality assurance.

The working capacity of the individual team members is shown in Table 2.

	Year 1	**Year 2**	**Year 3**
Martin Modrák	0.5 FTE	0.5 FTE	0.5 FTE
Postdoc 1	0.2 FTE	0.2 FTE	0.2 FTE
Diploma student 1	600 hours (~0.3 FTE)	300 hours (~0.15 FTE)	-
Diploma student 2	-	300 hours (~0.15 FTE)	600 hours (~0.3 FTE)

Table 2: Working capacity of the team members. FTE = full-time equivalent.

References

AKUTSU, T, MIYANO, S and KUHARA, S, 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. p. 17–28.

ARRIETA-ORTIZ, Mario L, HAFEMEISTER, Christoph, BATE, Ashley Rose, CHU, Timothy, GREENFIELD, Alex, SHUSTER, Bentley, BARRY, Samantha N, GALLITTO, Matthew, LIU, Brian, KACMARCZYK, Thadeous, SANTORIELLO, Francis, CHEN, Jie, RODRIGUES, C. D., SATO, Tsutomu, RUDNER, David Z, DRIKS, Adam, BONNEAU, Richard and EICHENBERGER, Patrick, 2015. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Molecular Systems Biology*. **11**(11), 839–839.

BERRONES, Arturo, JIMÉNEZ, Edgar, ALCORTA-GARCÍA, María Aracelia, ALMAGUER, F-Javier and PEÑA, Brenda, 2016. Parameter inference of general nonlinear dynamical models

of gene regulatory networks from small and noisy time series. *Neurocomputing*. **175**, 555–563.

BONNEAU, Richard, REISS, David J, SHANNON, Paul, FACCIOTTI, Marc, HOOD, Leroy, BALIGA, Nitin S and THORSSON, Vesteinn, 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*. **7**(5), 1.

BROOKS, Aaron N, REISS, David J, ALLARD, Antoine, WU, Wei-Ju, SALVANHA, Diego M, PLAISIER, Christopher L, CHANDRASEKARAN, Sriram, PAN, Min, KAUR, Amardeep and BALIGA, Nitin S, 2014. A system-level model for the microbial regulatory genome. *Molecular systems biology*. **10**(7), 740.

CARPENTER, Bob, GELMAN, Andrew, HOFFMAN, Matthew D., LEE, Daniel, GOODRICH, Ben, BETANCOURT, Michael, BRUBAKER, Marcus, GUO, Jiqiang, LI, Peter and RIDDELL, Allen, 2017. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*. **76**(1).

FU, Yao, JARBOE, Laura R and DICKERSON, Julie A, 2011. Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics*. **12**(1), 233.

GOUTSIAS, J and LEE, N H, 2007. Computational and experimental approaches for modeling gene regulatory networks. *Current pharmaceutical design*. **13**(14), 1415–1436.

GYGI, S P, ROCHON, Y, FRANZA, B R and AEBERSOLD, R, 1999. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*. **19**(3), 1720–30.

HUYNH-THU, V??n Anh, IRRTHUM, Alexandre, WEHENKEL, Louis and GEURTS, Pierre, 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*. **5**(9).

CHEBIL, I., NICOLLE, R., SANTINI, G., ROUVEIROL, C. and ELATI, M., 2014. Hybrid method inference for the construction of cooperative regulatory network in human. *IEEE Transactions on Nanobioscience*. **13**(2), 97–103.

CHEN, Yifei, LI, Yi, NARAYAN, Rajiv, SUBRAMANIAN, Aravind and XIE, Xiaohui, 2016. Gene expression inference with deep learning. *Bioinformatics*. **32**(12), 1832–1839.

IGLESIAS-MARTINEZ, Luis F., KOLCH, Walter and SANTRA, Tapesh, 2016. BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research. *Scientific Reports*. **6**(1), 37140.

LACHMANN, Alexander, GIORGI, Federico M., LOPEZ, Gonzalo and CALIFANO, Andrea, 2016. ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*. **32**(14).

MAIER, Tobias, GÜELL, Marc and SERRANO, Luis, 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Letters*. **583**(24), 3966–3973.

MALL, Raghvendra, CERULO, Luigi, GAROFANO, Luciano, FRATTINI, Veronique, KUNJI, Khalid, BENSMAIL, Halima, SABEDOT, Thais S, NOUSHMEHR, Houtan, LASORELLA, Anna, IAVARONE, Antonio and CECCARELLI, Michele, 2018. RGBM: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucleic Acids Research*.

MARGOLIN, Adam A, NEMENMAN, Ilya, BASSO, Katia, WIGGINS, Chris, STOLOVITZKY, Gustavo, DALLA FAVERA, Riccardo and CALIFANO, Andrea, 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*. **7 Suppl 1**(1), S7.

MODRÁK, Martin and VOHRADSKÝ, Jiří, 2018. Genexpi: A toolset for identifying regulons and validating gene regulatory networks using time-course expression data. *BMC Bioinformatics*. **Accepted.**

MORRISSEY, E.R., JUÁREZ, M.A., DENBY, K.J. and BURROUGHS, N.J., 2010. Inferring the topology of a non-linear sparse gene regulatory network using fully Bayesian spline regression. *Systems Biology*. p. 1–24.

RAMANIUK, Olga, ČERNÝ, Martin, KRÁSNÝ, Libor and VOHRADSKÝ, Jiří, 2017. Kinetic modeling and meta-analysis of B. subtilis sigA regulatory network during spore germination and outgrowth. *BBA Gene Regulatory Mechanisms*. **1860**(8), 894–904.

REISS, DJ, BALIGA, NS and BONNEAU, R, 2006. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC bioinformatics*.

TITSIAS, Michalis K, HONKELA, Antti, LAWRENCE, Neil D and RATTRAY, Magnus, 2012. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Systems Biology*. **6**(1), 53.

VEITIA, Reiner A, 2003. A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. *Biological reviews of the Cambridge Philosophical Society*. **78**(1), 149–70.

VOHRADSKY, Jiri, 2012. Stochastic simulation for the inference of transcriptional control network of yeast cyclins genes. *Nucleic Acids Research*. **40**(15), 7096–7103.

VOHRADSKÝ, Jiří, 2001. Neural Model of the Genetic Network. *Journal of Biological Chemistry*. **276**(39), 36168–36173.

VU, Tra Thi and VOHRADSKY, Jiri, 2007. Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae. *Nucleic Acids Research*. **35**(1), 279–287.

YEUNG, Ka Yee, DOMBEK, Kenneth M., LO, Kenneth, MITTLER, John E., ZHU, Jun, SCHADT, Eric E., BUMGARNER, Roger E. and RAFTERY, Adrian E., 2011. Construction of regulatory networks using expression time-series data of a genotyped population. *Proceedings of the National Academy of Sciences*. **108**(48), 19436–19441.

ZHANG, X., LIU, K., LIU, Z.-P., DUVAL, B., RICHER, J.-M., ZHAO, X.-M., HAO, J.-K. and CHEN, L., 2013. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*. **29**(1), 106–113.

ZOPPOLI, Pietro, MORGANELLA, Sandro and CECCARELLI, Michele, 2010. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC bioinformatics*. **11**(1), 154.