Information-theoretic analysis of multivariate single - cell signaling responses using SLEMI

Tomasz Jetka[1], **Tomasz Winarski**[1], **Karol Nienałtowski**[1], **Sławomir Błoński**[1], **and Michał Komorowski**[1,*]

[1]Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland
[*]m.komorowski@sysbiosig.org

ABSTRACT

Mathematical methods of information theory constitute essential tools to describe how stimuli are encoded in activities of signaling effectors. Exploring the information-theoretic perspective, however, remains conceptually, experimentally and computationally challenging. Specifically, existing computational tools enable efficient analysis of relatively simple systems, usually with one input and output only. Moreover, their robust and readily applicable implementations are missing. Here, we propose a novel algorithm to analyze signaling data within the framework of information theory. In contrast to existing techniques, our approach enables statistically and computationally efficient and robust analysis of signaling systems with high-dimensional outputs and a large number of input values. Analysis of the NF-κB single - cell signaling responses to TNF-α uniquely reveals that the NF-κB signaling dynamics improves discrimination of high concentrations of TNF-α with a modest impact on discrimination of low concentrations. Our readily applicable R-package, SLEMI - statistical learning based estimation of mutual information, allows the approach to be used by computational biologists with only elementary knowledge of information theory.

Availability:

(i) *Supplemental Information* - containes theoretical and experimental methods:
https://github.com/sysbiosig/SLEMI/blob/master/paper/SI.pdf

(ii) R package SLEMI
http://github.com/sysbiosig/SLEMI

(iii) *User Manual* - contains the documentation of the package
https://github.com/sysbiosig/SLEMI/blob/master/paper/Manual.pdf

(iv) *Testing procedures* - contain step-by-step instructions to assist with the package's installation and running essential functions
https://github.com/sysbiosig/SLEMI/blob/master/paper/TestingProcedures.pdf

1 Introduction

Biochemical descriptions of cellular signaling require quantitative support to explain how complex stimuli (inputs) are translated and encoded in distinct activities of pathway's effectors (outputs)[1]. Information theory and probabilistic modeling offer an attractive approach. Regardless of specific details of a signaling pathway, within information theory, a signaling system can be considered as an input-output device that measures an input signal, x, by eliciting a stochastic output, Y. In a typical example, the input, x, is the concentration of a ligand that activates a receptor, and the output, Y, is the activity of a signaling effector, which might be the nuclear concentration of an activated transcription factor quantified over time[2–4]. As signaling systems are inherently stochastic, the input-output relationship is usually represented by the probability distribution $P(Y|X=x)$. The overall fidelity of signaling systems is within information theory summarised by the information capacity, C^*. The information capacity is expressed in bits, and generally speaking, 2^{C^*} represents the maximal number of different inputs that a system can effectively resolve (e.g., different ligand concentrations)[5,6]. The interest in the unique perspective of information theory is increasing with broader availability of single-cell data. However, exploring the information-theoretic perspective experimentally remains conceptually and technically challenging. Moreover, existing computational tools are computationally and statistically inefficient, and their applicability to study systems with multiple inputs and outputs is limited. Also, we lack readily applicable implementations. Here, we propose a novel algorithm, that is computationally efficient, provides accurate estimates for relatively small sample size, and, hence, can provide novel biological insight for systems with highly-dimensional outputs and a large number of inputs. We also provide the algorithm's robust implementation.

In a typical experiment aimed to quantify how much information can flow through a given signaling system, input values $x_1 \leq x_2 ... \leq x_m$, ranging from 0 to saturation are considered. Then, responses to each input level, x_i, are quantified in a large number of individual cells. Responses of individual cells are represented as vectors y^i_j, where j varies from 1 to the number of

quantified cells, denoted as n_i. Often, vectors y_j^i contain entries that quantify activities of signaling effectors over time. The data are assumed to follow an unknown distribution, $y_j^i \sim P(Y|X=x_i)$, which represents the system's input-output relationship. To estimate the information capacity, existing algorithms[4,7,8] utilise the data, y_j^i, to construct approximations, $\hat{P}(Y|X=x_i)$, of the output distributions, $P(Y|X=x_i)$, for i ranging from 1 to the number of input values m. Thereafter, the approximations $\hat{P}(Y|X=x_i)$ are used to evaluate the mutual information (MI)

$$MI(X,Y) \approx \sum_{i=1}^{m} \int_{\mathbb{R}^k} \hat{P}(y|X=x_i)P(x_i)\log_2 \frac{\hat{P}(y|X=x_i)}{\hat{P}(y)} dy, \tag{1}$$

where $P(x_i)$ is a distribution of input values, which is usually set depending on the context, and $\hat{P}(y) = \sum_{i=1}^{m} \hat{P}(y|X=x_i)P(x_i)$ is the approximation to the marginal distribution of the output, Y. The maximization of MI with respect to usually unknown probabilities of input values $P(x_i)$ allows computation of the information capacity

$$C^* = \max_{P(x_1),\ldots,P(x_m)} MI(X,Y). \tag{2}$$

The available algorithms differ in the way, in which, the approximation of the output distributions, $\hat{P}(y|X=x_j)$, is constructed. Specifically, Blahut - Arimoto (BA) algorithm[7–10] uses a discrete approximation. All possible values of responses are divided into a finite set of intervals and frequencies of responses falling into the same interval as y_j^i are used as the approximation $\hat{P}(y_j^i|X=x_j)$. On the other hand, the approach of[4], following earlier work[11], uses the k-nearest neighbors (KNN) method, in which continuous approximations $P(y_j^i|x_i)$ are constructed based on the distance of y_j^i to the k-th most similar response. Each of the above approaches is limited by the dimensionality of the output, Y. Practically, the BA algorithm can be applied to systems with the one-dimensional output. On the other hand, for multidimensional output, an accurate estimation of $P(Y|X=x_i)$ using KNN requires a large sample size.[12]. Moreover, KNN demands arbitrary specification of the parameter k, which for insufficient data size does not generally guarantee unbiased estimation[11–14], and yields estimation sensitive to arbitrary assumptions. Moreover, KNN based approaches often require solving computationally expensive optimisation problems. In Section 1 of the *Supplemental Information* (*SI*), we provide more background on information theory and existing computational tools. Here further, we introduce an alternative framework that allows efficient, in terms of sample size and computational time, estimation of the information capacity for systems with high dimensional outputs, Y. In addition, our approach uniquely provides probabilities of correct discriminations between different input values. The framework reveals that NF-κB signaling dynamics improves discrimination of high concentrations of TNF-α with a modest impact on discrimination of low concentrations. A robust implementation of the proposed computational tools is also provided.

2 Results

2.1 Efficient computation of the information capacity

In contrast to existing approaches, instead of estimating high dimensional conditional output distributions $P(Y|X=x_i)$, we propose to estimate the discrete, conditional input distribution, $P(x_i|Y=y)$, which is known to be a simpler problem[15]. Estimator of $P(x_i|Y=y)$, denoted here further as $\hat{P}(x_i|Y=y)$, can be built by using bayesian classification methods, here specifically, logistic regression. Estimation of the MI using $\hat{P}(x_i|Y=y)$ rather than $\hat{P}(y|X=x_i)$ is possible as the MI (Eq. 1), can be alternatively written as[5]

$$MI(X,Y) \approx \sum_{i=1}^{m} P(x_i) \sum_{j=1}^{n_j} P(y_j^i|X=x_i)\log_2 \frac{\hat{P}(x_i|Y=y_j^i)}{P(x_i)}. \tag{3}$$

Therefore, for a given $P(x_i)$ and $\hat{P}(x_i|Y=y)$ MI can be evaluated without knowledge of $P(Y|X=x_i)$. Although $P(Y|X=x_i)$ is still present in the above sum, it represents averaging of the term $\log_2 \frac{\hat{P}(x_i|Y=y_j^i)}{P(x_i)}$ over available data, which can be achieved with data y_j^i alone, without explicit knowledge of $P(y|X=x_i)$. Further, the above formulation allows to employ an efficient convex optimization scheme to compute C^* from MI. Therefore, no numerical gradient optimization is needed. In Section 2 of the *SI* we describe the algorithm in detail and we prove its mathematical correctness. In Box S1 we present the algorithm as pseudocode.

The logistic regression used to approximate $\hat{P}(x_i|Y=y)$ combined with the convex optimization led to the algorithm that outcompetes existing approaches in terms of sample size need to provide accurate estimates, computational time, and robustness to algorithm settings. As opposed to the KNN based approaches[4,11], algorithm's settings do not impact the estimates, which ensures robust estimation. The approach is also an order of magnitude faster compared to KNN method, especially for systems with high number of input values. These advantages are demonstrated in Section 3 of the *SI*, specifically in Figures S1 and

S2. The benefits are of particular importance for signaling systems with multidimensional outputs, Y, and a large number of considered input values, x_i.

Importantly, the algorithm, also, uniquely allows analyzing signaling systems in terms of discrimination error. Precisely, the information capacity *per se* does not tell us, which input values cells can effectively distinguish. It only provides an overall measure of signaling fidelity. Given that our approach is based on the approximation of the conditional input distribution, $\hat{P}(x_i|y)$, the probabilities of correct discrimination are readily available. It can be shown[5] that the strategy that maximizes the probability of correct guessing, which input x_i lead to observed output y, is the maximum *a posteriori* rule, which selects x_i with highest $\hat{P}(x_i|Y=y)$. We describe the calculation of the probabilities of correct discrimination in Section 4.5 of the *SI*.

The advantages of the proposed framework extend beyond statistical and computational aspects. To demonstrate this, we have experimentally studied, single - cell signaling responses of the NF-κB system to a range of concentrations of TNF-α in murine embryonic fibroblasts cell line. Analysis of the experimental data revealed how information transfer is distributed over time. It also uniquely showed that the dynamics of the NF-κB signaling responses leads to improved discrimination of high TNF-α concentrations with a minor effect on discrimination of low concentrations.

2.2 Signaling dynamics of NF-κB system strongly improves discrimination of only high TNF-α concentrations

The NF-κB signaling is one of the key pathways involved in the control of the immune system. It is also one of the first cellular signaling systems studied within the framework of information theory. So far, several papers quantified its information capacity, e.g.,[4,8,16]. Interestingly, response dynamics have been shown to have greater signaling capacity compared to non-dynamic responses[4,16]. To demonstrate what novel insight can be gained with our framework, we have measured NF-κB responses (y^i_j's in the above notation) to a range of 5 minutes pulses of TNF-α concentrations (x_i's), in single - cells, using life confocal imaging. Experimental methods are described in Sections 4.1 - 4.3 of the *SI*. Fig. 1A shows temporally resolved responses, y^i_j, to representative four concentrations, whereas Fig. S4 to all considered concentrations. In order to provide a further insight into the dynamic aspect of signaling, we used the data to calculate the information capacity for two different scenarios: time-point and time-series responses. For time-point responses, we consider experimental measurements at a specified time only. On the other hand, for time-series responses, we consider measurements from the beginning of the experiment till an indicated time. Fig. 1B and C show information capacity for time-point and time-series responses as a function of time. Time-series data include time-point data, which implies higher information content. Precisely, information capacity for time-series responses increases sharply over 1 bit at $\sim$ 25 min., and reaches $\sim$1.3 bits at late times, i.e., $\sim$120 minutes. In contrast, information capacity for time-point responses reaches 1 bit around the time of maximal response, i.e., $\sim$20 min, only, and remains below 1 bit for all other times. Interestingly, time-point responses exhibit a second peak of information transfer at $\sim$85 minutes. This is an extension of the result of[4] and[16], where signaling dynamics, represented by time-series, have been shown to increases the information capacity. The efficiency of the algorithm, uniquely, allowed to calculate the capacity as the function of time, and, hence, reveal how the information transfer is distributed over time. These computations involved outputs containing up to 40 entries, which is usually not achievable with other approaches. Most importantly, however, with our approach, we can decipher what information is transferred using the additional ~ 0.3 bits provided by response dynamics. To address this, we have calculated the probabilities of correct discrimination between all concentration pairs. This can be done within our framework as it is based on the estimation of the conditional input distribution, $\hat{P}(x_i|Y=y)$. For each single - cell response y^i_j, we found most likely input value and compared whether the true one is the one most likely. When most likely value matched the correct one we interpreted this as the correct discrimination. Calculated probabilities of correct pairwise discrimination for time-point and time-series responses are presented as pie-charts in Fig.1D and Fig.1E, respectively. Random discrimination yields 0.5 chance of correct guessing. Hence, all probabilities are ≥ 0.5. Comparison of Fig.1D and Fig. 1E, demonstrates that time-point and time-series responses result with similar probabilities of correct discrimination between low and high concentrations, i.e., pie-charts close to full circle. On the other hand, discrimination between high concentrations is largely improved for time-series responses. For instance, discrimination between 0 and 100 ng/ml is close to perfect for both scenarios. On the other hand, discrimination between 8 and 100 ng/ml based on time-point responses is close to random, whereas it is more than 75% successfully based on time-series responses. This demonstrates that signaling dynamics contains information that improves discrimination of high TNF-α concentrations, which is uniquely revealed by our computational approach. In the light of this analysis, the higher capacity of time-series responses results largely from improved discriminability of high concentrations rather than improved discriminability of all concentrations. Sections 4.3 - 4.5 of the *SI* contain more details on the analysis of experimental data.

2.3 R-package

Our algorithm is available as robustly implemented R-Package SLEMI. It can be used by a computational biologist with a limited background in information theory. Details on installation and applicability are provided in the package's *User Manual*.

Step-by-step *Testing Procedures* are also provided to assist with package's installation and running essential functions. The package includes the NF-κB dataset as well as scripts to reproduce Fig. 1. Computations needed to plot each panel of the figure, without bootstrap, do not exceed several minutes on a regular laptop.

3 Summary

Compared to existing approaches, our framework significantly simplifies information-theoretic analysis of cellular signaling systems. It benefits from a novel algorithm, which is based on the estimation of the discrete input distribution as opposed to the estimation of continuous output distributions in conventional approaches. Conveniently, the algorithm does not involve numerical gradient optimization. Our approach results not only in short computational times but, most importantly, in relatively low sample sizes needed to obtain accurate estimates. Therefore, our framework enables efficient analysis of signaling data with a large number of inputs and high dimensional outputs. Analysis of multidimensional data is increasingly relevant with a broader availability of multivariate measurement techniques. Importantly, the approach relates the information capacity to the probability of discrimination between different input values. Last but not least, we provide the first software package for information theoretic analysis of single - cell signaling data that is readily applicable by a user with only elementary knowledge of information theory.

Acknowledgements

The immortalised murine embryonic fibroblasts cell line (3T3) expressing fluorescent fusion proteins relA-dsRed was kindly provided by prof. Savas Tay. The focus on the NF-κB pathway was inspired by prof. Tomasz Lipniacki. Experimental component of this research was carried out with the use of CePT infrastructure financed by the European Regional Development Fund within the Operational Program *Innovative Economy* for 2007-2013. TJ was supported by his own funds and the European Commission Research Executive Agency under grant CIG PCIG12-GA-2012- 334298, TW by IUVENTUS PLUS grant IP2012016572, MK by the Polish National Science Centre under grant 2015/17/B/NZ2/03692.

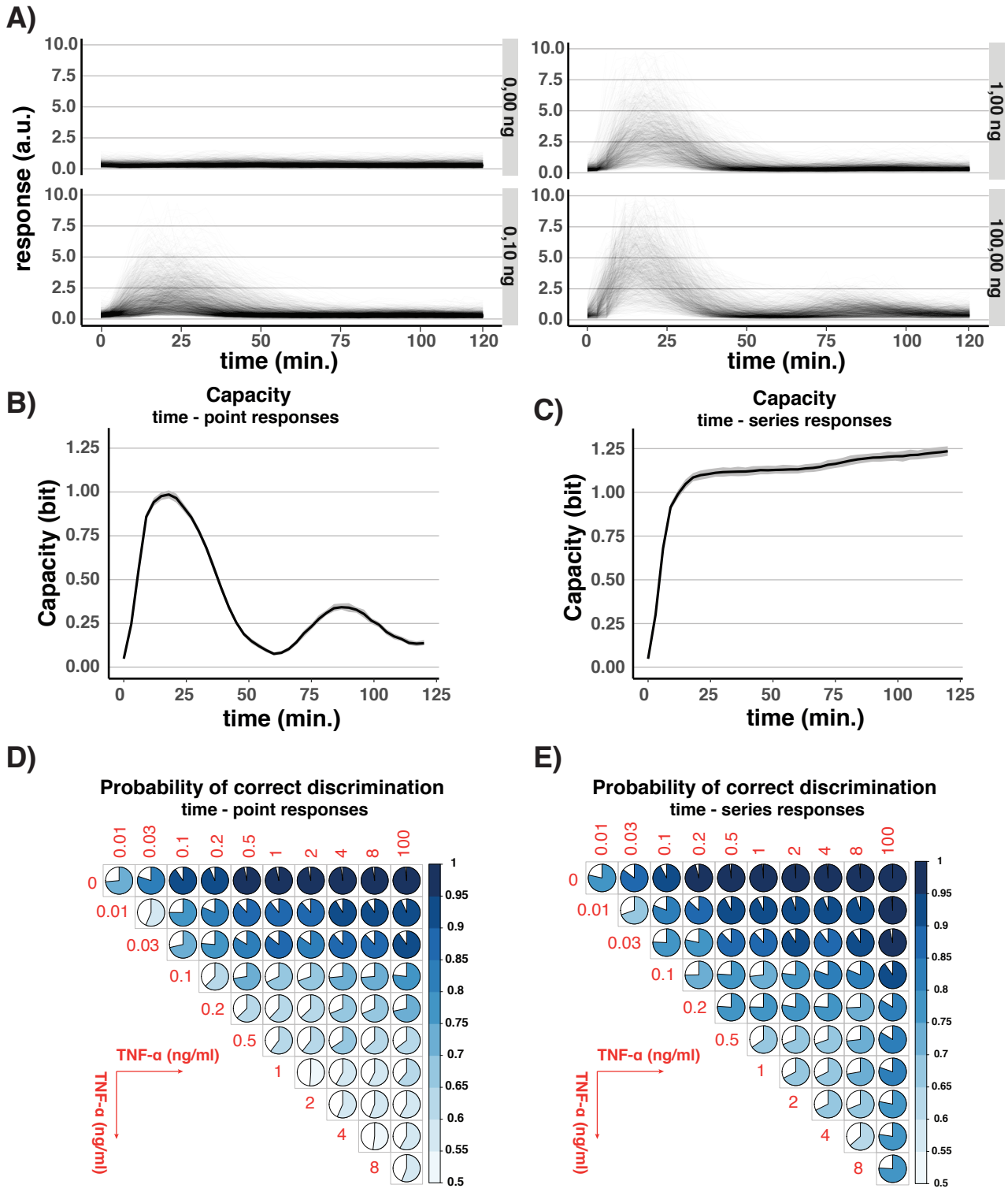

Figure 1. Information-theoretic analysis of the NF-κB responses to TNF-α stimulation. **(A)** Temporally resolved responses of individual cells to selected concentrations of TNF-α. The panel corresponds to Fig S4. **(B)** Information capacity as a function of time for time-point responses (see text). **(C)** As in (B) but for time-series responses. **(D)** Probabilities of the correct discrimination in pairwise discrimination between TNF-α concentrations for time-point responses. The color filled fraction of the circle marks the probability of correct discrimination (see text and *SI*). **(E)** The same as in (D) but for time-series responses. Modeling details: Uncertainties of estimates (grey ribbons in B and C) were obtained by bootstrapping 80% of data (repeated 100 times). Probabilities in D and E present mean of 50 bootstrap re-sampling.

References

1. Purvis, J. E. & Lahav, G. Encoding and decoding cellular information through signaling dynamics. *Cell* **152**, 945–956 (2013).
2. Bowsher, C. G. & Swain, P. S. Identifying sources of variation and the flow of information in biochemical networks. *Proc. Natl. Acad. Sci.* **109**, E1320–E1328 (2012).
3. Levchenko, A. & Nemenman, I. Cellular noise and information transmission. *Curr. opinion biotechnology* **28**, 156–164 (2014).
4. Selimkhanov, J. *et al.* Accurate information transmission through dynamic biochemical signaling networks. *Science* **346**, 1370–1373 (2014).
5. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
6. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).
7. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Inf. Theory* **18**, 460–473 (1972).
8. Cheong, R., Rhee, A., Wang, C. J., Nemenman, I. & Levchenko, A. Information Transduction Capacity of Noisy Biochemical Signaling Networks. *Science* **334**, 354–358 (2011).
9. Vontobel, P. O. A generalized Blahut- Arimoto algorithm. In *Information Theory, 2003. Proceedings. IEEE International Symposium on*, 53 (IEEE, 2003).
10. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Inf. Theory* **18**, 14–20 (1972).
11. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. review E* **69**, 066138 (2004).
12. Mack, Y. & Rosenblatt, M. Multivariate k-nearest neighbor density estimates. *J. Multivar. Analysis* **9**, 1–15 (1979).
13. Wang, Q., Kulkarni, S. R. & Verdú, S. Divergence estimation for multidimensional densities via *k*-nearest-neighbor distances. *IEEE Transactions on Inf. Theory* **55**, 2392–2405 (2009).
14. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci.* 201309933 (2014).
15. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, NY, USA:, 2001).
16. Zhang, Q. *et al.* Nf-κb dynamics discriminate between tnf doses in single cells. *Cell systems* **5**, 638–645 (2017).