Applying Molecular Communication Theory to Estimate Information Loss in Cell Signal Transduction: An Approach Based on Cancer Transcriptomics

Zahmeeth Sakkaff
Department of Computer Science and Engineering
University of Nebraska-Lincoln
zsayedsa@cse.unl.edu

Aditya Immaneni
Department of Computer Science and Engineering
University of Nebraska-Lincoln
aimmaneni@cse.unl.edu

Massimiliano Pierobon
Department of Computer Science and Engineering
University of Nebraska-Lincoln
pierobon@cse.unl.edu

ABSTRACT

Signal transduction pathways are chemical communication channels embedded in biological cells, and they propagate information from the environment to regulate cell growth and proliferation, among other cell's behaviors. Disruptions in the normal functionalities of these channels, mostly resulting from mutations in the underlying genetic code, can be leading causes of diseases, such as cancer. Motivated by the increasing availability of public data on genetic code expression in cell tissue samples, *i.e.*, transcriptomics, and the emerging field of molecular communication, a novel data-driven approach based on experimental data mining and communication theory is proposed in this paper. This approach is an alternative to existing computational models of these pathways in the context of cancer, which often appear to oversimplify the complexity of the underlying mechanisms. In contrast, a computational methodology is here derived to estimate the difference in information propagation performance of signal transduction pathways in healthy and diseased cells, solely based on transcriptomic data. This methodology is built upon a molecular communication abstraction of information flow through the pathway and its correlation with the expression of the underlying DNA genes. Numerical results are presented for a case study based on the JAK-STAT pathway in kidney cancer, and correlated with the occurrence of pathway gene mutations in the available data.

KEYWORDS

Molecular Communication, Mutual information, Signal Transduction Pathways, The Cancer Genome Atlas (TCGA), Genomics, Transcriptomics

1 INTRODUCTION

Biological cells have the natural ability to sense information from the environment through the reception, propagation, and processing of molecular signals. This ability is at the basis of major cellular functionalities, such as the regulation of the cell growth rate and proliferation [22]. Dysregulations in normal functionalities of these cellular information systems, *i.e.*, signaling pathways, are considered at the basis of complex diseases such as cancer [30], which is nowadays one of the leading causes of death worldwide. One of the obstacles to effective strategies for early diagnosis and cure of these diseases is that the molecular mechanisms leading to signaling pathway dysregulation, which have their roots in alterations of the underlying genetic code (mutations), are largely ambiguous [29]. While mainstream high-throughput analysis techniques, such as Next Generation Sequencing (NGS) [21], are providing an increasingly massive amount of data on multi-level molecular profiles of healthy and diseased tissues [28], advanced integrative and interdisciplinary strategies are required to abstract meaningful information on the molecular bases of these diseases.

Molecular communication theory is an emerging transdisciplinary research field devoted to the modeling, characterization, and engineering of communication systems where information propagates through molecules and chemical reactions [1]. This field is experiencing an increasing interest in the study of biological cells [26], and their communication functionalities [11], with the overarching goal of realizing future programmable biological devices, enabled by the latest advances in synthetic biology [19], pervasively interconnecting biological systems with the Internet, *i.e.*, the Internet of Bio-Nano Things [2].

In this paper, we propose the idea that the aforementioned diseases originating from altered cellular information processing can be successfully studied through a transformative data-driven approach that draws tools from molecular communication theory. Existing cancer signaling pathway models, mostly based on either boolean logic or Ordinary Differential Equations (ODEs) [4, 8, 20], tend to mechanistically capture in different levels of complexity the chemical processes at the basis of the propagation of the signal from the extracellular signaling molecules, *e.g.*, growth factors, to the downstream gene regulation, while the effect of genetic mutations on the disruption of these mechanisms is mostly reduced to very simplistic assumptions [10, 17]. As an alternative to the complexity of realistic computational models of these processes, this paper introduces a novel perspective where information and communication theory, already proven effective for cancer classification and analysis [13], are integrated with cancer high-throughput data for the quantitative study and characterization of the disruption in the flow of molecular information through signaling pathways caused by genetic changes.

NANOCOM '18, September 5–7, 2018, Reykjavik, Iceland, 2018

ACM ISBN 978-1-4503-5711-1/18/09…$15.00
https://doi.org/10.1145/3233188.3233202

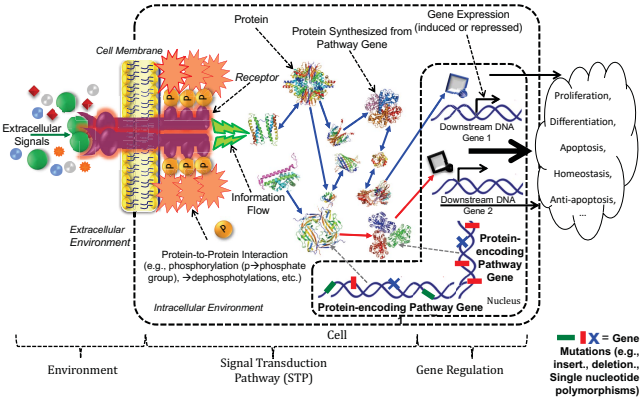

Figure 1: Pictorial sketch of the biochemical processes in cell signal transduction pathways and pathway genes.

Our approach is based on a molecular-communication-theoretic abstraction of the genetic processes underlying signal transduction pathways, in particular centered around the transcription of genetic information, which ultimately controls the performance of the pathway in the propagation of information. This approach is motivated by the increasing wealth of public data based on transcriptomics technologies [28], which provide measurements on the transcription of the cell's genes from experimental samples. The latter are in fact characterized by high efficiency and low cost with respect to other technologies that directly measure the interacting molecules (proteins) that propagate information through a pathway, *i.e.*, proteomics. Based on the aforementioned abstraction, we derive a computational methodology to estimate the difference in information propagation performance of signal transduction pathways in healthy and diseased cells solely based on transcriptomic data, and we correlate these results with the most common pathway gene mutations underlying diseases.

The rest of the paper is organized as follows. In Sec. 2, we review the biochemistry underlying signal transduction pathways and their genetic underpinnings, and propose our abstraction, in Sec. 3 we review the types of data handled in our work, and propose a methodology to estimate the communication performance of a pathway from transcriptomic data. In Sec. 4 we present a case study based on the well-known JAK-STAT pathway and its relevance in kidney cancer. Finally, in Sec. 5 we conclude the paper.

2 MOLECULAR COMMUNICATION ABSTRACTION OF INFORMATION LOSS IN CELL SIGNAL TRANSDUCTION

2.1 The Biochemistry of Signal Transduction Pathways and Pathway Genes

In this paper, we focus on the biochemical processes that enable the propagation of information from the external environment through biological cells, *i.e.*, signal transduction pathways, where it ultimately controls numerous cell behaviors, and the role played by their genetic underpinnings, *i.e.*, the pathway genes [22].

As shown in Fig. 1, these processes are based on chemical interactions between specific biological macromolecules, or proteins, which form cascades of information-propagating mutual activation reactions called phosphorylations. In each phosphorylation reaction, a specific protein, the kinase, acquires a phosphate group through a reaction with another activated kinase, therefore becoming activated, and, possibly after binding to other proteins into complexes, can subsequently activate another protein, *e.g.*, another kinase, downstream of the cascade. The activated kinases are then "reset" into a non-activated state by removal of the phosphate group (dephosphorylation), operated by other specific proteins, the phosphatases. These chained reactions are initiated by special proteins, the **receptors**, usually located across the **cell membrane**, which modulate the phosphorylation of other proteins in the intracellular environment downstream of the cascade according to **extracellular signals**, *i.e.*, values of physical or chemical parameters, such as hormones. This modulation propagates the information contained in the extracellular signals through the cascaded phosphorylations, until controlling the activation of other special proteins, the **transcription factors**, which in turn regulate the expression of one or more **downstream DNA genes** inside the cell **nucleus** [9]. This downstream DNA gene regulation usually results in the activation (amplification) or repression (attenuation) of a specific cell behavior, such as its growth rate/division, *i.e.*, **proliferation**, a specific cell characteristic, *i.e.*, **differentiation**, the probability of inducing death, *i.e.*, **apoptosis**, **anti-apoptosis**, and physiological stability, *i.e.*, **homeostasis**, among others.

The aforementioned proteins involved in the signal transduction pathway are present in the cell at determinate concentrations, which ensure a "healthy" propagation of information through the cell that maintains the adaptability of the organism to the extracellular environment [22]. These protein molecules are inevitably subject to degradation, or simply digested by the cell (proteolysis), as at the same time new ones are continuously synthesized from DNA genes, *i.e.*, **pathway genes**, as depicted in Fig. 1. A pathway gene is a stretch of DNA that codes for the sequence of amino acids that composes a specific pathway protein, which is synthesized from the gene through the fundamental processes of transcription and translation. Transcription is initiated by the enzyme (a specific type of protein) RNA polymerase (RNAP) that binds to the promoter region of the gene, starting the production of the messenger RNA (mRNA) molecules. These latter molecules carry the genetic information of the protein encoded in the gene to the ribosome, the protein assembly machinery. Subsequently, ribosomes, which are able to recognize and bind to the mRNA molecules, complete the synthesis of the corresponding protein through the process of translation, by assembling together the component amino acids as instructed by the mRNA. As with the aforementioned downstream DNA genes, also the pathway genes may be subject to regulation as a result of the information propagated by the pathway, therefore increasing or decreasing their transcription rate through activation or repression, respectively. As a result, the concentrations of mRNAs present in the cell, which are measured through transcriptomics [28], such as those at the basis of the TCGA database considered in this paper, may be correlated to the information propagated through the pathway from the extracellular signals.

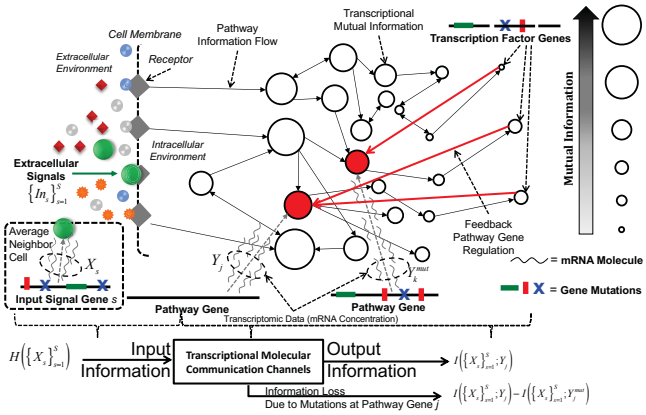

Figure 2: Proposed molecular communication abstraction of information loss caused by pathway gene mutations and based on transcriptomics.

DNA genes in the cell can be subject to **mutations**, *i.e.*, permanent changes in their code sequence resulting from error during DNA duplication in cell division, or from direct damages of its structure by external causes [22]. Some of these changes in the aforementioned pathway genes can result into changes in pathway protein structure and function, with consequent alteration of the protein behavior in terms of propagation of information within the cell signaling pathway, thus leading to characteristic features of tumor cells, such as uncontrolled proliferation, anti-apoptosis (cell uncontrolled survival), or unbalance in cell differentiation signals leading to metastases [27]. DNA mutations might appear in different forms, but in the preliminary work described in this paper we will consider for simplicity only gene mutations, *i.e.*, where there is a change in nucleotide sequence within a gene, as follows: **insertion** of additional nucleotides into the sequence, **deletion** of existing nucleotides, substitution of a nucleotide with a different nucleotide (**Single Nucleotide Polymorphism - SNP**), at specific positions of the DNA sequence of the gene, as graphically depicted in Fig. 1. Gene mutations are generally more frequent in cancer than other types of mutations, *i.e.*, chromosome mutations [14], although the relevance of this latter type to gene expression in cancer has been recently studied [3]. In this paper, we abstract and characterize the correlation between pathway gene mutations and the loss in information as it propagates through cell signal transduction pathways by utilizing tools from communication and information theory applied to genomic and transcriptomic data.

2.2 Molecular Communication Abstraction Based on Transcriptomics

According to the reference molecular communication abstraction proposed in this paper, a cell signaling pathway is modeled as a network of molecular communication channels, which propagate **input information** carried by the extracellular signals through the pathway, where **information loss** may occur due to the aforementioned possible **gene mutations**, until being relayed as **output information** to the transcription factors that regulate the downstream genes. Our goal is to estimate the amount of input and output information, as well as the information loss associated to gene mutations leading to diseases, from the knowledge of the pathway reaction map [15] and publicly available transcriptomic data [28]. With reference to Fig. 2, we base our estimations on the following assumptions:

- *We estimate the information that propagates through the pathway by taking into account the concentrations of mRNA molecules transcribed from specific protein-encoding DNA genes, instead of the proteins themselves.* These concentrations, called transcripts, are the data obtained by transcriptomic technologies applied to experimental cell samples [21]. Although information is in fact propagated in the pathway by their corresponding synthesized proteins through their interactions and mutual activation, as described in Sec. 2.1, and our assumption is supported by recent studies on the correlation between transcripts and the corresponding proteins [6]. The latter work demonstrated experimentally that, for each specific gene, the relation between the steady-state concentration of transcripts, obtained through transcriptomics, and the corresponding number of synthesized proteins can be approximated in human cells by a constant gene-specific mRNA-protein ratio.
- *Although cell signaling pathways are dynamic, and their input-output behavior is in general function of the time, we restrict our analysis to the observed average output of the pathway that corresponds to each possible observed input.* This is motivated by the fact that the aforementioned transcriptomic data are obtained as average samples from an ensemble of non-synchronized signaling pathways, *i.e.*, an average of the states of all the cells included in each experimental tissue sample [28].
- *For the signal transduction pathway under analysis, we rely on the knowledge of the pathway reaction map [15], namely, the set of all the chemical reactions in the pathway, to understand the relationship between the pathway genes.* Each pathway reaction expressed in general as $A + B \rightleftharpoons C + D$, where A, B are the reactant molecule species, C, D are the product molecule species, respectively (for irreversible reactions the backward arrow is omitted, and B and/or D can be omitted depending on the reaction), is translated into edges as shown in Fig. 2. In particular, *the red edges abstract the regulation of pathway gene expressions operated by the aforementioned transcription factors*, whose activity is controlled by the information that propagates in the pathway.
- *We consider the presence of the aforementioned types of mutations at each gene if they appear in the genomic data of diseased samples but not in healthy samples.* These will enable the study of how specific gene mutations correlate to changes in the information that propagates through the signal transduction pathway in case of disease. The classification of each cell sample into healthy or having a specific diseased, *i.e.*, cancer, is given a priori by the aforementioned public databases.

As a consequence of these assumptions, we can estimate the **Input Information** from the values of the transcripts X_s for each **Input Signal Gene** $s = 1, \ldots, S$, which are the genes that encode the extracellular signals. We can consider these as the average transcripts of all the cells in each tissue sample, and therefore proportional to the average extracellular signals that are input of the pathway In_s

from the neighboring cells in the tissue. We quantify this information through the entropy expression $H\left(\{X_s\}_{s=1}^{S}\right)$ (which we call **Transcriptional Entropy**), which is related to the information carried by the extracellular signals In_s as follows:

$$H\left(\{X_s\}_{s=1}^{S}\right) = H\left(\{In_s\}_{s=1}^{S}\right) + \sum_{s=1}^{S} \log_2 K_{in,s}\,, \qquad (1)$$

where $K_{in,s}$ is the aforementioned gene-specific mRNA-protein ratio for each input signal gene s, and the formula corresponds to the information entropy of an ensemble of scaled continuous random variables [5]. The information $H\left(\{In_s\}_{s=1}^{S}\right)$ carried by the extracellular signals propagates through the pathway by modulating the interactions between the pathway proteins, until affecting the activity of the transcription factors downstream of the pathway, as described in Sec. 2.1. In turn, these transcription factors regulate the expression of a set of pathway genes (gene j and gene k in red in Fig. 2). Biological noise and other effects [16] tend to decrease the information content in the protein interaction modulation by randomization or equivocation [5] during its propagation in the signaling pathway, resulting in a residual **Output Information at each pathway gene** j that can be quantified through the **Transcriptional Mutual Information** $I = I\left(\{X_s\}_{s=1}^{S}; Y_j\right)$, where Y_j represents the transcripts of pathway gene j. For the latter, we can express the following:

$$I\left(\{X_s\}_{s=1}^{S}; Y_j\right) = H\left(\{X_s\}_{s=1}^{S}\right) - H\left(\{X_s\}_{s=1}^{S} | Y_j\right) = I\left(\{In_s\}_{s=1}^{S}; Y_j\right)\,, \qquad (2)$$

since the conditional transcriptional entropy $H\left(\{X_s\}_{s=1}^{S} | Y_j\right)$, which is computed over values of the transcripts X_s, is equal to the conditional entropy $H\left(\{In_s\}_{s=1}^{S} | Y_j\right)$ plus the same factor as in (1). When the mutual information is computed, the two factors cancel out, leading to (2). The **Information Loss due to Mutations at the pathway gene** j is computed as the difference between the transcriptional mutual information $I_j = I\left(\{X_s\}_{s=1}^{S}; Y_j\right)$ computed from transcriptomic samples classified as healthy, and the same but computed from transcriptomic samples classified as diseased, which includes the effect of mutations on the propagation of information, denoted as $I_j^{mut} = I\left(\{X_s\}_{s=1}^{S}; Y_j^{mut}\right)$. Finally, the occurrence, or frequency, of mutations in diseased samples with respect to healthy samples can be considered and classified on the basis of the information loss results. In this paper, we will limit the latter to a visual comparison of the information losses at specific genes and the observed mutation frequencies, while more structured approaches will be tackled in future work.

3 ESTIMATING TRANSCRIPTIONAL MUTUAL INFORMATION

3.1 Data

We obtain the data necessary to estimate the aforementioned transcriptional information parameters according to the following steps:

- We select a **specific cell signal transduction pathway** (or a set of cross-communicating pathways) and retrieve the corresponding pathway reaction map. For this, we interrogate the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database, which is an integrated database resource for biological interpretation of gene sequences and high throughput genomic data [12, 15]. Along with the pathway maps, the KEGG Pathway Database provides the list of the corresponding pathway genes, together with other information on their exact sequence, other specific molecular interactions, regulation reactions, and crosstalk with other signal transduction pathways.
- We select a **specific disease**, *i.e.*, a cancer type, and retrieve the transcriptomic data corresponding to the pathway genes listed by the KEGG Pathway Database. For this, we interrogate The Cancer Genome Atlas (TCGA) [28], a publicly available high-throughput genomic database. In particular, the provided transcripts have been obtained from different tissue specimens of healthy and diseased patients through the RNA-Seq-based transcriptome sequencing technique, and their corresponding concentration values are expressed in units of Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Proteins within the same functional family are represented in the network map with the same node, but they expressed by different genes. In the scope of this paper, we sum all the transcripts related to these proteins before utilizing them in to estimate transcriptional mutual informations.

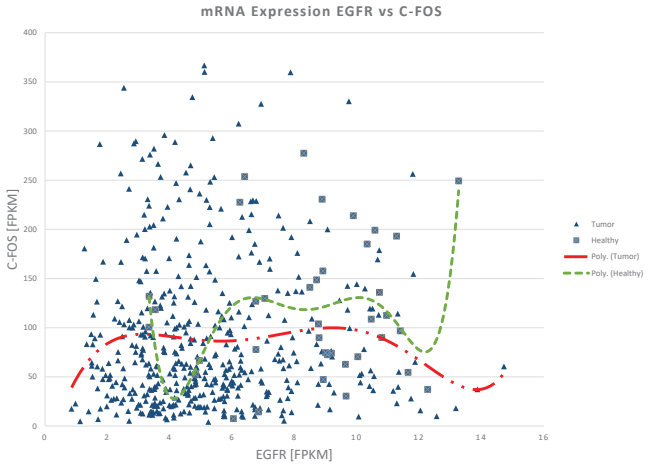

Figure 3: Transcripts of EGFR and c-Fos genes from TCGA obtained from healthy and tumor tissue samples in subjects affected by colorectal cancer.

As an example, in Fig. 3 we show data points extracted from TCGA colorectal cancer data for the EGF/EGFR pathway in *Homo sapiens*. For each sample, we extracted the transcripts of the EGFR input signal gene and the corresponding transcripts of the pathway gene c-Fos [23], that is regulated by the pathway transcription factors, as explained in Sec. 2.2. By observing the polynomial trend lines [7] of the healthy samples Vs the tumor sample, it is clear that in healthy cells the expression of c-Fos can be controlled more precisely and over a larger range than in tumor cells, where c-Fos is mostly around a stable (and low, as also confirmed in [18]) value for almost all range of observed EGFR expression.

3.2 Computational Methodology

The final goal of our computational methodology is the estimation of the transcriptional mutual information $\tilde{I}_j$ at each pathway gene j, expressed as

$$\tilde{I}_j = \tilde{H}\left(\{X_s\}_{s=1}^{S}\right) - \tilde{H}\left(\{X_s\}_{s=1}^{S}|Y_j\right), \tag{3}$$

where $\tilde{H}(.)$ and $\tilde{H}(.|.)$ denote the estimated joint entropy and conditional entropy, respectively, X_s is the value of the transcripts for input signal gene s and Y_j is the value of the transcripts for pathway gene j (or functional family j).

The *estimated input entropy* $\tilde{H}\left(\{X_s\}_{s=1}^{S}\right)$ is computed through the histogram approach [24] as

$$\tilde{H}\left(\{X_s\}_{s=1}^{S}\right) = -\sum_{i_1=1}^{N_{b_1}} \cdots \sum_{i_S=1}^{N_{b_S}} \prod_{i=1}^{S} p_{X_s}\left(x_{s,i_s}\right) \cdot \sum_{i=1}^{S} \log_2\left(\frac{p_{X_s}\left(x_{s,i_s}\right)}{w_{X_s}}\right), \tag{4}$$

where $p_{\{X_s\}_{s=1}^{S}}\left(\{x_{s,i_s}\}_{s=1}^{S}\right)$ is the probability for the value of X_s corresponding to the i_s-th histogram bin x_{s,i_s}, and N_{b_s} and w_{X_s} are the number and size of histogram bins considered to approximate the probability density function of X_s according to the available transcriptomic data for the input signal gene s, respectively. The formulation of the expression in (4) is based on the simplifying assumption of having input signals with independent probability distributions.

The *estimated conditional entropy* $\tilde{H}\left(\{X_s\}_{s=1}^{S}|Y_j\right)$ of the input signal gene transcripts $\{X_s\}_{s=1}^{S}$ given the transcripts of the pathway gene j (or functional family j) is computed as

$$\tilde{H}\left(\{X_s\}_{s=1}^{S}|Y_j\right) = -\sum_{h=1}^{N_{b_{Y_j}}} p_{Y_j}(y_{j,h}) \cdot \sum_{i_1=1}^{N_{j,h,b_1}} \cdots \sum_{i_S=1}^{N_{j,h,b_S}} \prod_{i=1}^{S} p_{X_s|Y_j}\left(x_{s,i_s}|y_{j,h}\right) \cdot \sum_{i=1}^{S} \log_2\left(\frac{p_{X_s|Y_j}\left(x_{s,i_s}|y_{j,h}\right)}{w_{X_s|y_{j,h}}}\right), \tag{5}$$

where $N_{b_{Y_j}}$ is the number of bins considered to approximate the probability density function of Y_j according to the available transcriptomic data for the pathway gene j, $p_{X_s|Y_j}\left(x_{s,i_s}|y_{j,h}\right)$, N_{j,b_s} and $w_{X_s|y_{j,h}}$ are the probability for the value of X_s corresponding to the i_s-th histogram bin x_{s,i_s}, the number and size of histogram bins, respectively, considered to approximate the probability density function of X_s according to the available transcriptomic data for the input signal gene s that is at the input when the gene j histogram bin value $y_{j,h}$ is considered as the signal transduction pathway output.

The numbers of histogram bins N_{b_s}, $N_{b_{Y_j}}$, and N_{j,h,b_s}, for $s = 1, \ldots, S$ and the h-th transcript value for the pathway gene j histogram are computed from the gene transcriptomic data according to the Doane's formula [24] as follows:

$$N_b = 1 + \log_2(C) + \log_2\left(1 + \frac{g_A}{\sigma_{g_A}}\right). \tag{6}$$

where C is the total number of available transcriptomic data samples, *i.e.*, number of healthy or diseased tissue samples, g_A is the estimated 3rd-moment-skewness of the transcript distribution p_A, and $\sigma_{g_A} = \sqrt{\frac{6(C-2)}{(C+1)(C+3)}}$. To obtain the values for N_{b_s}, $N_{b_{Y_j}}$, and N_{j,h,b_s}, the parameter A is substituted with X_s, Y_j, and $X_s|Y_j = y_{j,h}$, respectively. Finally, the histogram bin sizes w_{X_s} and $w_{X_s|y_{j,h}}$ are computed by dividing the difference between the maximum and minimum values of the transcripts X_s or $X_s|y_{j,h}$, respectively, by the corresponding number of histogram bins computed through (6).

For example, if we apply this computational method to estimate the transcriptional mutual information for the data shown in Fig. 3, where we consider only one input signal gene ($S = 1$), *i.e.*, EGFR, and the pathway gene c-Fos, we obtain $I_{c-Fos} = I\left(X_{EGFR}; Y_{c-Fos}\right) \approx 0.73$ bits (41 tissue samples) and $I^{mut}_{c-Fos} = I\left(X_{EGFR}; Y^{mut}_{c-Fos}\right) \approx 0.27$ bits (470 tissue samples) considering in each case only the data from the healthy samples and tumor samples, respectively. This corresponds to $I_{c-Fos} - I^{mut}_{c-Fos} = 0.46 bits$, which quantifies the information loss suffered by the EGF/EGFR pathway for tumor tissue cells in colorectal cancer on the mechanism of regulation of the c-Fos gene by the EGF signal.

4 CASE STUDY AND NUMERICAL RESULTS

In this section, we present numerical results obtained with the computational method presented in this paper to estimate transcriptional mutual information, and information loss due to pathway gene mutations leading to disease. In this case study, we focus on a specific signaling pathway, *i.e.*, the JAK-STAT pathway. This pathway is a crucial signaling cascade for several extracellular signals, within the functional families of cytokines, hormones and growth factors. When JAK is activated, it stimulates cell growth, differentiation, migration and apoptosis [22]. These factors are crucial for the normal functioning of biological processes, such as immune system, hematopoiesis, lactation, and development of adipocytes. Any mutation that affects the activity of JAK regulation signaling can cause diseases, such as inflammation, leukemia, erythrocytosis [25]. In particular, in our study, we seek to understand how the normal regulation and dysregulation of JAK-STAT is associated to cancer (kidney cancer, as detailed in the folloting), hence this study could be provide novel insights to the cancer research community. Moreover, the JAK-STAT pathway is known for its simplicity with respect to other signaling pathways in eukaryotic cell.

By following the methodology described in Sec. 3, we retrieved the standard *Homo sapiens* JAK-STAT signaling pathway reaction map from the KEGG Pathway Database, and used this model to extract all the genes associated with each process along the pathway. As a result, we have a total of 161 genes involved in the JAK-STAT pathway, which can be grouped into 32 functional families, as defined in Sec. 3.1, including the aforementioned 3 functional families of input signal genes, including 46 cytokines, 10 hormones and, 3 growth factors. Next, we extracted the transcripts for all the 161 genes from the TCGA to apply our computational method and obtain the transcriptional mutual information of the pathway. Within

the TCGA, we chose to sample kidney tissues cells as our case study because we were able to find a greater amount of gene expression data associated to the pathway under analysis. Finally, we retrieved transcriptomic data for the healthy (128 healthy tissue samples) and diseased (893 tumor tissue samples) cases.

Similarly, we extracted mutation data for all 161 genes considering three different commonly known mutation types as follows: i) insertion, which involves the addition of one or more nucleotide base pairs into a DNA sequence, ii) deletion, which is when a part of a DNA sequence is lost during DNA replication, and iii) SNP, which comprises a variation that occurs in a single nucleotide (T, C, G, or A).

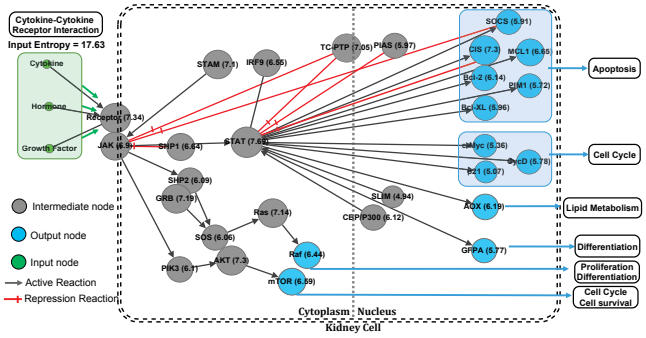

Figure 4: Transcriptional mutual information for JAK-STAT pathway genes in healthy kidney cells.

For the estimation of the input entropy, the transcriptomic data of healthy and diseased tissues are combined to utilize the entire range of transcript values that the input signal genes can take. The transcriptional mutual information values for each gene (which includes the corresponding funcitonal family) in the KEGG JAK-STAT pathway is estimated as described in Sec. 3.2. When utilizing the healthy 128 tissue samples, transcriptional mutual information values are reported in Fig. 4, and graphically shown in a corresponding proportional size of each pathway map node. As expected, the transcriptional mutual information values are decreasing as the information propagates through the pathway, accumulating chemical noise at each reaction (data processing inequality), from an estimated input joint entropy $\tilde{H}(X) = 17.63$ bits to estimated multiple outputs of MI SOCS $\tilde{I}_j = 5.91$ bits, CIS $\tilde{I}_j = 7.3$ bits, MIC $\tilde{I}_j = 6.65$ bits, BcI $\tilde{I}_j = 6.14$ bits, PIM1 $\tilde{I}_j = 5.72$ bits, BcI-XL $\tilde{I}_j = 5.96$ bits, c-Myc $\tilde{I}_j = 5.36$ bits, CycD $\tilde{I}_j = 5.78$ bits, p21 $\tilde{I}_j = 5.07$ bits, AOX $\tilde{I}_j = 6.19$ bits, GFPA $\tilde{I}_j = 5.77$ bits, Raf $\tilde{I}_j = 6.44$ bits, and mTOR $\tilde{I}_j = 6.59$ bits.

Similarly, in Fig. 5 we report transcriptional mutual information values of each gene of the pathway when considering the diseased 893 tissue samples. We observe the same behavior in the trend of the mutual information values along the pathway, albeit these values are overall lower than those observed in the healthy case, thus confirming the impairments in the propagation of information along the pathway caused by disease-leading gene mutations.

In Fig. 6 we show the difference in the transcriptional mutual information values between the healthy and diseased cases, which correspond to quantifying the information loss that affects the

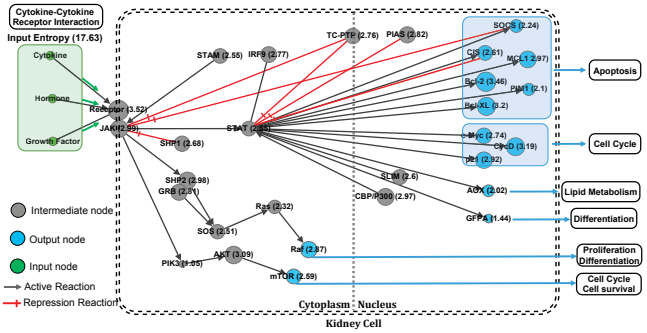

Figure 5: Transcriptional mutual information for JAK-STAT pathway genes affected by mutations in diseased (tumor) kidney cells.

JAK-STAT pathway in the presence of gene mutations. Here, we computed the information loss in percentage for each gene in the pathway by considering the fraction of mutual information loss between the healthy and disease cases with respect to the mutual information of the healthy case. Here we notice that the largest loss of information occurs at the STAT gene. This is expected since this pathway gene is (negatively) regulated by the largest number of transcription factors (red edges), and, as a consequence, its transcripts will carry most of the information propagating through the pathway, as explained in Sec. 2.2. When the information propagating in the pathway is affected by greater impairments caused by gene mutations, the transcriptional information of this particular pathway gene will suffer from larger losses.

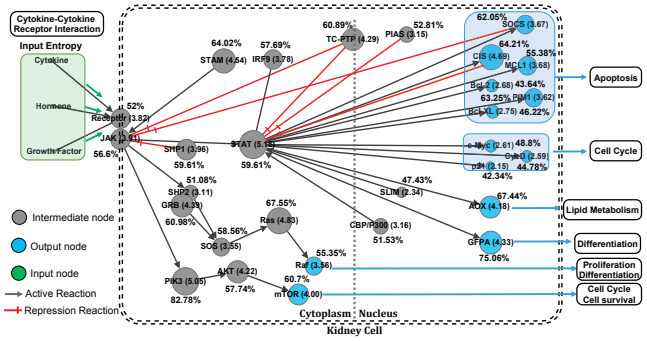

Figure 6: Transcriptional mutual information loss between healthy and mutated JAK-STAT pathway genes in healthy and tumor kidney cells, respectively.

Finally, we looked at the mutation data to correlate the aforementioned information losses with the frequency of insertion, deletion, and SNP mutations. By considering a total of 291 tissue samples from the TCGA and 161 genes listed the JAK-STAT pathway model from the KEGG Pathway Database, we filtered out a total of 1203 variant type mutations, which correspond to 28 insertions, 68 deletions and 1107 SNPs. In Fig. 7, we show a stacked bar chart to compare the occurrence of the three types of mutations in each of the gene functional families. The Receptor family of genes in the

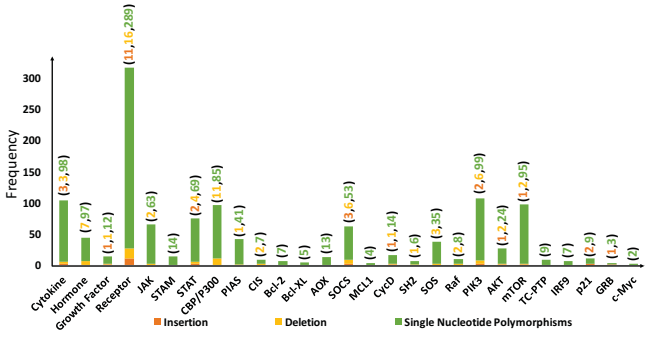

Figure 7: Comparison of mutation frequencies for insertion, deletion and SNP for JAK-STAT pathway genes in diseased (tumor) kidney cells.

JAK-STAT pathway shows the highest likelihood of mutation with 11 insertions, 16 deletions and 289 SNP mutations over the entire data set. As mentioned above, the effect of this high frequency of mutation is seen in the downstream regulated gene STAT, which experiences the highest loss of information.

5 CONCLUSIONS

In this paper, we proposed the idea that diseases originating from altered cellular information processing can be successfully studied through a transformative data-driven approach that draws tools from molecular communication theory and is grounded on available experimental data, rather than computational models. In this direction, we defined a molecular-communication-theoretic abstraction of the genetic processes underlying signal transduction pathways in biological cells, in particular centered around the transcription of genetic information, which ultimately controls the performance of the pathway in the propagation of information. Based on this abstraction, we derived a computational methodology to estimate the difference in information propagation performance of signal transduction pathways in healthy and diseased cells solely based on publicly available transcriptomic data, and we correlated these results with the most common pathway gene mutations underlying diseases. We finally provided proof-of-concept numerical results for a case study based on the JAK-STAT pathway in kidney cancer, and correlated with the occurrence of pathway gene mutations in the available data. We believe that this approach will set the basis for further research in a novel direction for communication theory, and yet provide a *novel tool for cancer research to characterize diseases through a standard methodology and information-theoretic metrics.* Further research is envisioned in the integration of this methodology with *ad hoc* design of experimental platforms to obtain further data, *e.g.*, through proteomics.

ACKNOWLEDGMENTS

This work was supported by the NIH National Institutes of General Medical Sciences through grant 5P20GM113126-02, and the US National Science Foundation through grant CCF-1816969.

REFERENCES

[1] I. F. Akyildiz, J. M. Jornet, and M. Pierobon. Nanonetworks: A new frontier in communications. *Communications of the ACMs*, 54(11):84–89, November 2011.

[2] I. F. Akyildiz, M. Pierobon, S. Balasubramaniam, and Y. Koucheryavy. The internet of bio-nano things. *IEEE Communications Magazine*, 53(3):32–40, March 2015.

[3] B. Alaei-Mahabadi, J. Bhadury, J. W. Karlsson, J. A. Nilsson, and E. Larsson. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *PNAS*, 113(48):13768–13773, 2016.

[4] J. Bachmann, A. Raue, M. Schilling, V. Becker, J. Timmer, and U. Klingmuller. Predictive mathematical models of cancer signalling pathways. *J Intern Med*, 271: 155–165, 2012.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd Edition.* Wiley, 2006.

[6] F. Edfors, F. Danielsson, B. M. Hallström, L. Käll, E. Lundberg, F. Pontén, B. Forsström, and M. Uhlén. Gene-specific correlation of rna and protein levels in human cells and tissues. *Mol Syst Biol.*, 12(10):883, 2016.

[7] J. Fan. Local polynomial modelling and its applications: From linear regression to nonlinear regression. *Monographs on Statistics and Applied Probability. Chapman & Hall/CRC*, pages 1269–1275, 1996.

[8] J. Fisher and T. Henzinger. Executable cell biology. *Nat Biotechnol.*, 25(11): 1239–49, November 2007.

[9] E. Gonçalves, J. Bucher, A. Ryll, J. Niklas, K. Mauch, S. Klamt, M. Rocha, and J. Saez-Rodriguez. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. *Molecular BioSystems*, 9 (7):1576–1583, 2013.

[10] I. Habibi, E. S. Emamian, and A. Abdi. Quantitative analysis of intracellular communication and signaling errors in signaling networks. *BMC Systems Biology*, 8(89):1–15, 2014.

[11] C. Harper, M. Pierobon, and M. Magarini. Estimating information exchange performance of engineered cell-to-cell molecular communications: a computational approach. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*, April 2018.

[12] D. A. Harrison. The jak/stat pathway. *Cold Spring Harbor perspectives in biology*, 4(3):a011205, 2012.

[13] W.-C. Hsu, C.-C. Liu, F. Chang, and S.-S. Chen. Cancer classification: Mutual information, target network and strategies of therapy. *Journal of Clinical Bioinformatics*, 2(16):1–11, 2012.

[14] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502: 333–339, 2013.

[15] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40 (D1):D109–D114, 2011.

[16] J. E. Ladbury and S. T. Arold. Noise in cellular signaling pathways: causes and effects. *Trends Biochem Sci.*, 37(5):173–178, May 2012.

[17] R. Layek, A. Datta, M. Bittner, and E. R. Dougherty. Cancer therapy design based on pathway logic. *BIOINFORMATICS*, 27(4):548–555, December 2010.

[18] S. Mahner, C. Baasch, J. Schwarz, S. Hein, L. Wölber, F. Jänicke, and K. Milde-Langosch. C-fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma. *Br. J. Cancer*, 99(8):1269–1275, October 2008.

[19] A. Marcone, M. Pierobon, and M. Magarini. A parity check analog decoder for molecular communication based on biological circuits. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*, May 2017.

[20] W. Materi and D. Wishart. Computational systems biology in drug discovery and development: methods and applications. *Drug Discov Today*, 12(7-8):295–303, April 2007.

[21] M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet.*, 11 (1):31–46, January 2010.

[22] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*, chapter 12.2, pages 425–429. W. H. Freeman, 2005.

[23] M. K. Pandey, G. Liu, T. K. Cooper, and K. M. Mulder. Knockdown of c-fos suppresses the growth of human colon carcinoma cells in athymic mice. *Int J Cancer*, 130(1):213–222, January 2012.

[24] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes, 4th ed.* McGraw-Hill, 2002.

[25] J. S. Rawlings, K. M. Rosler, and D. A. Harrison. The jak/stat signaling pathway. *Journal of cell science*, 117(8):1281–1283, 2004.

[26] Z. Sakkaff, J. L. Catlett, M. Cashman, M. Pierobon, N. R. Buan, M. B. Cohen, and C. A. Kelley. End-to-end molecular communication channels in cell metabolism: an information theoretic study. In *Proc. of the ACM International Conference on Nanoscale Computing and Communication (ACM NanoCom)*, September 2017.

[27] R. Sever and J. S. Brugge. Signal transduction in cancer. In *Cold Spring Harb Perspect Med.*, volume 5, page a006098, 2015.

[28] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[29] B. Vogelstein, N. Papadopoulos, V. Velculescu, S. Zhou, L. Diaz, and K. K. Jr. Cancer genome landscapes. *Science*, 339(6127):1546–58, March 2013.

[30] D. H. R. Weinberg. Hallmarks of cancer: the next generation. *Cell.*, 144(5):646–74, March 2011.