Seminars in Cell & Developmental Biology 51 (2016) 14–23

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

# Information theory in systems biology. Part II: protein–protein interaction and signaling networks

Zaynab Mousavian[a], José Díaz[b], Ali Masoudi-Nejad[a,*]

[a] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Grupo de Biología Teórica y Computacional, Centro de Investigación en Dinámica Celular, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico

ARTICLE INFO

*Article history:*
Received 7 December 2015
Accepted 7 December 2015
Available online 12 December 2015

*Keywords:*
Protein–protein interaction networks
Signaling networks
Entropy
Mutual information
Channel capacity and rate distortion theory

ABSTRACT

By the development of information theory in 1948 by Claude Shannon to address the problems in the field of data storage and data communication over (noisy) communication channel, it has been successfully applied in many other research areas such as bioinformatics and systems biology. In this manuscript, we attempt to review some of the existing literatures in systems biology, which are using the information theory measures in their calculations. As we have reviewed most of the existing information-theoretic methods in gene regulatory and metabolic networks in the first part of the review, so in the second part of our study, the application of information theory in other types of biological networks including protein–protein interaction and signaling networks will be surveyed.



**Contents**

1. Introduction ..... 14
2. Preliminaries from information theory ..... 15
   2.1. Discrete channel and channel capacity ..... 15
   2.2. Rate distortion theory ..... 15
3. Applications of information theory in systems biology ..... 15
   3.1. Protein–protein interaction network ..... 16
      3.1.1. Finding protein complexes ..... 16
      3.1.2. Complexity analysis ..... 17
      3.1.3. Identification of biomarkers ..... 17
      3.1.4. Study of network robustness ..... 18
   3.2. Signaling networks ..... 18
4. Conclusion ..... 22
   References ..... 23

## 1. Introduction

Information theory was developed by Claude Shannon, to primarily address the main limitations on signal processing operations, including data compression and reliable data storing and communication. After the beginning of information theory, it has rapidly widened to find applications in many other research topics such as probability theory, statistics, mathematics, economics, computer science, physics, and also to more away disciplines such as bioinformatics and systems biology. Systems biology is the study of biological components at the system level, and information theory provides a theoretical framework to study the relations between components.

In this review, we address the wide applications of information theory in the field of systems biology. Only a limited reviews has been published which have focused on a particular biological network like signaling network [1–5] or a specific

* Corresponding author.
*E-mail addresses:* zmousavian@ut.ac.ir (Z. Mousavian), amasoudin@ibb.ut.ac.ir (A. Masoudi-Nejad).
*URL:* http://LBB.ut.ac.ir (A. Masoudi-Nejad).

http://dx.doi.org/10.1016/j.semcdb.2015.12.006
1084-9521/

application such as reverse engineering of cellular networks [6,7], and they do not attempt to cover all applications of information theory in all biological networks. As in the first part of our review, gene regulatory networks and metabolic networks have been surveyed, so in the second part, the main focus puts on the information-theoretic studies in the area of protein–protein interaction networks and signaling networks. In protein–protein interaction network, various problems including complex identification, complexity analysis, determining subnetwork markers, among others, have been addressed by the use of information theory. Furthermore, to study the rate of signal transmission in signaling networks and estimate the channel capacity in such networks, information theoretic approaches have been published by researchers.

The rest of this manuscript is organized as follows: Section 2 briefly introduces some basic concepts of information theory framework, for more details readers can refer to the first part of this review. Moreover, two new information-theoretic measures like the channel capacity and the rate distortion theory will be introduced in Section 2. Different applications of information theory in protein–protein interaction and signaling networks will be surveyed in Section 3. Finally, the second part of our review will be concluded in Section 4.

## 2. Preliminaries from information theory

The preliminary concepts of information theory have been introduced in the first part of our review. However, we have a brief introduction about the basic information theoretic concepts in this part and the readers can refer to the first part of our review for more details. Furthermore, channel capacity and rate distortion theory are two important information-theoretic concepts which have been used in signaling networks, and we have also an introduction about them in this section.

A fundamental concept of information is *entropy* which characterizes the amount of uncertainty for prediction of the value of a random variable. The *entropy* of a discrete random variable $X$ with alphabet $\chi$ and the probability mass function $p(x)$ is defined as:

$$H(X) = -\sum_{x \in \chi} p(x) \cdot \log\, p(x)$$

Let $X$ and $Y$ be two discrete random variables having the joint probability mass function $p(x,y)$, the *joint entropy* is calculated as:

$$H(X, Y) = -\sum_{x \in \chi} \sum_{y \in Y} p(x, y) \cdot \log\, p(x, y)$$

The entropy of a random variable conditional on the knowledge of another random variable is called *conditional entropy* and calculated as follows:

$$H(Y|X) = \sum_{x \in \chi} p(x) H(Y|X = x)$$

A relation can be defined between the *joint entropy* and the *conditional entropy* as:

$$H(X, Y) = H(X) + H(Y|X)$$

Another type of entropy is the *relative entropy*, also known as *Kullback–Leibler Distance*, which quantifies the distance of two distribution functions. The *relative entropy* of two distributions with the probability functions $p(x)$ and $q(x)$ is defined as:

$$D(p||q) = \sum_{x \in \chi} p(x) \cdot \log \frac{p(x)}{q(x)}$$

*Mutual information* is a key measure of information theory and it can be assumed as a *relative entropy* between the joint distribution $p(x,y)$ and the product distribution $p(x)\cdot p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x \in \chi} \sum_{y \in \Upsilon} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} \\ &= D(p(x, y)||p(x).p(y)) \end{aligned}$$

The *mutual information* also indicates the size of shared information between two random variables. So it can measure the amount of reduction in the uncertainty about one random variable when another variable is known:

$$I(X; Y) = H(X) - H(X|Y)$$

Finally, the *conditional mutual information* is defined as the shared information between two random variables due to the knowledge of the third variable:

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z)$$

### 2.1. Discrete channel and channel capacity

A *discrete channel* can be defined as a system containing channel input and output symbols $\mathcal{A}$ and $\mathcal{B}$ respectively. The probability of having the symbol $b$ in output given that the symbol $a$ has been sent in channel input is represented in the probability transition matrix $(b|a)$.

A channel is called a memoryless channel, if the probability distribution of the output is not conditionally dependent on the previous channel inputs or outputs and only depends on the channel input at the current time.

Now we can define the *information channel capacity* $C$ as:

$$C = \max_{p(x)} (I(X; Y))$$

in which $X$ and $Y$ are defined over $\mathcal{A}$ and $\mathcal{B}$ respectively. For a discrete memoryless channel, $p(x)$ is the distribution of input random variable $x$ and the maximum of $I$ is taken over all possible distributions $p(x)$.

Some properties are associated with the *channel capacity* measure:

- $C \geq 0$
- $C \leq \log |\mathcal{A}|$ and $C \leq \log |\mathcal{B}|$

### 2.2. Rate distortion theory

The distance between the random variable $X$ and its image $\hat{X}$ is called a *distortion* measure. For example the number of bits which is required for describing a real number is infinite and therefore a finite representation of it can never be perfect. The *rate distortion theory* addresses the problem of finding the minimal number of bits per each source symbol that should be communicated over channel such that the source can be approximately reconstructed at the receiver without exceeding a predefined amount of distortion $D$. It can be formally defined as a following minimization problem:

$$R(D) = \min_{E[d(X,\hat{X})] \leq D} I(X, \hat{X})$$

## 3. Applications of information theory in systems biology

Systems biology is an emerging discipline that aims to understand complex biological systems by computational and mathematical modeling, and information theory can help to achieve this. In following, different studies will be reviewed which