Objective Distances for Transcriptome Analyses
According to the Thermodynamic Model

Tomokazu Konishi
konishi@akita-pu.ac.jp

Bioresource Sciences, Akita Prefectural University, Shimoshinjyo, Akita 010-0195, Japan

Keywords: thermodynamics, transcriptome, microarray, distance, clustering

1 Introduction

Data analysis and interpretation require intellectual frameworks, and to share the results in a community, scientific bases are preferred for the framework. A variety of indexes have been used for interpreting distances between transcriptome data, however, such indexes always lack necessity since they have been derived for non-biological purposes and their appropriateness have not been verified for transcriptome analyses. Here introduced are a definition of distance and two indexes that are derived from thermodynamic model of transcriptome formation [3, 4], which describes transcriptome by way of interactions between protein factors and nucleic acids. For example, a human cell may have 1,000 species of protein factors, and its transcriptome can be described by the same numbers of dimensions of Gibbs free energies and Arrhenius activation energies. Consequently, distance between any two expression levels can be found as Euclidean distance calculated with the energies. However, unfortunately, such data treatment is presently impossible, since the constant relationship between a protein factor and a gene is not available yet.

2 Method and Results

2-1 1-norm distance between gene expressions

This determination method of distance describes those between genes and experiments for clustering analyses [2]. In methods frequently used, the distance between expression levels is defined by Pearson correlation, which does not indicate length but similarity. In contrast, here it is explained by the energies given by the protein factors. Since the energies presently cannot be found but they can be represented by the logratios of expression levels, which is determined by microarray analyses [4]. Similarly, the distance between experiments can be found as the arithmetic mean of the distances of genes. Since the dimensions of the energies are 1,000 as far, each logratio data should not be treated as independent vectors; they will affect too much if Euclidean distance is applied. Rather, they are likely to be treated to have a same dimension, especially if the numbers of genes are closely focused. As a result, these ideas recommend one of the methods that have been prepared [1]; i.e. city-block distance with average linkage for both genes and experiments. However, the ideas bring an unit for the distance, the logratio (Fig. 1), as a linear norm for the distance. With the unit, any two tree graphs can be compared as like maps with proper scale indicated.

The vertical bar indicated at the left most side shows distance of ratio 1.10.

Figure 1: Sample of hierarchical clustering of experiments using the 1-norm distance and average linkage.
sympathy index and opposition index

The indexes are introduced to find genes or experiments related to particular gene or experiment, providing aids for finding distances with paying attention to some particular protein factors. For example, when activity concentration of a regulator increases, chances of the factor bind to certain regulatory sequences of genes will increase. Since each the binding site has constant spatial relationship with a promoter and RNA-polymerase II, the bound regulator increases or decreases the expression levels according to the relationship. Consequently, the same regulator can increase some genes and decrease other genes simultaneously, and, those oppositely affected genes may have reverse functions. This idea is plausible by means of functions; genes with opposed functions should be regulated in opposed ways. Additionally, although close distance calculated by the 1-norm method suggests similar functions of genes or treatment in experiments, such relationships will become evident by omitting affections by no-related genes. The calculation of the opposition (sympathy) index from one gene to other genes are performed as for gene of series of experiments \( i (i = 1, 2, 3, \ldots n) \) with \( m \) genes data, the index between subjected experiment \( s \) is calculated by logratios between particular control sample as

1. let \( \text{score}_{i, g} = 0 \), if \( \text{logratio}_{i, g} \times \text{logratio}_{s, g} > 0 \) \((-\infty \) when sympathy index is calculated \)
2. let \( \text{score}_{i, g} = |\text{logratio}_{i, g}| \), if \( |\text{logratio}_{i, g}| < |\text{logratio}_{s, g}| \)
3. let \( \text{score}_{i, g} = |\text{logratio}_{s, g}| \), if \( |\text{logratio}_{i, g}| \geq |\text{logratio}_{s, g}| \)
4. let \( \text{index}_{s, i} = 1/m \left\{ \sum_{g=1}^{m} (\text{score}_{i, g}) \right\} \)

The algorithm can be used to find the index for distance in gene expressions with subjected gene \( s \) and for various genes \( i \) of experiment \( g \) with \( m \) experiments.

The opposition and sympathy indexes are independent and can be visualized with a 2D-plot with logratios as units. In a sample plot (Fig. 2), a group of experiments that determines effects of a same treatment to cultured cells that possess different constructs of RNAi was shown. The subjected experiment to be compared was without RNAi, which is presented as experiment A in the Fig. 1. Among the experiments, the one with the least relationship is experiment K, which is the mock control. In experiment M, many expression levels follow exp. A, while some genes that seem to be in the downstream of the interfered gene are oppositely regulated. Those indexes will help to find hidden relationships as well.

Figure 2: A sample of opposition-sympathy plot

References


