Ordering Genes by Significance of Occurrences

Jun Sese[1] sesejun@gi.k.u-tokyo.ac.jp
Shinichi Morishita[1] moris@k.u-tokyo.ac.jp
Kousaku Okubo[2] kousaku@imcb.osaka-u.ac.jp

[1] Department of Complexity Science and Engineering, Graduate School of Frontier Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2] Institute for Molecular and Cellular Biology, Osaka University, 1-3 Yamadaoka, Suita 565-0871, Japan

Keywords: significance, expression profile, mRNA, dbEST, bodymap

1 Introduction

Genes differentially expressed in different tissues, during development, or during specific pathologies are of foremost interest to both basic and pharmaceutical research. Digital expression profile of non-normalized mRNAs like Bodymap [2, 3] is a way of finding such genes. Although the number of mRNAs registered in dbEST is numerous, it is hard to identify the specific library where one mRNA is expressed, because the mRNA maybe obtain from a normalization of various libraries. We therefore focus on non-normalized mRNA library and we establish an index of finding differentially expressed genes. Then we might express that the genes ordering by using our statistical method in the libraries is useful to reduce the number of cDNAs to put on micro arrays.

2 Method

2.1 Comparison Between Two Expression Profiles

Audic and Claverie [1] propose an statistical method for evaluating "differential display" that compares two expression profiles.

Let us consider the situation of average occurrence x out of N samples in one tissue. We pick up new N samples and the probability of occurrence y is according to Poisson distribution; namely, $\frac{e^{-x}x^y}{y!}$.

In the case of two tissues, for instance T_1 and T_2, we have to consider influence of the random fractuation and the difference of the number of samples in each tissue. Suppose that we observe x occurrences of gene G out of N_1 occurrences of all genes in tissue T_1, while we find y out of N_2 in tissue T_2.

tissue	T_1	T_2
occurrences of all genes	N_1	N_2
occurrences of G	x	y

The probability of occurrences of G in T_2 on the basis of T_1 is:

$$\begin{aligned}\Pr(T_2|T_1) &= \int_0^\infty d\lambda_2 \int_0^\infty d\lambda_1 p(d=\lambda_1|x)p(y|d=\lambda_2)\delta\left(\lambda_2 - \frac{N_2}{N_1}\lambda_1\right) \\ &= \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}\end{aligned}$$

Lower value of $\Pr(T_2|T_1)$ indicates that the significance of y occurrence in T_2 is higher.

2.2 Expansion to Multiple Expression Profiles

The main of proposal of this paper is to expand the concept of differential display to more than two tissues and to define a statistical measurement. We define the samples and occurrence of a gene G in a tissue as following table.

tissue	T_1	T_2	T_3	$\dots$	T_n
occurence of all genes	N_1	N_2	N_3	$\dots$	N_n
occurrence of G	x_1	x_2	x_3	$\dots$	x_n

Expansion1: Probability of occurrences of a gene G in T_1 **and** T_2 on the basis of other tissues.

The probability of occurrences of a gene G in tissue T_1 based on tissues $T_3, \dots, T_n$ is independence of that in tissue T_2.

$$\Pr(T_1 \cap T_2 | T_3 \cup \ldots \cup T_n) = \Pr(T_1 | T_3 \cup \ldots \cup T_n) \times \Pr(T_2 | T_3 \cup \ldots \cup T_n)$$

Expansion2: Probability of occurrences of a gene G in T_1 **or** T_2 on the basis of other tissues.

Suppose that new tissue T_{12} has the number of occurrences $x_1 + x_2$, the number of samples $N_1 + N_2$. The occurrence in tissue T_1 is individual of that in tissue T_2.

$$\Pr(T_1 \cup T_2 | T_3 \cup ... \cup T_n) = \Pr(T_1 | T_3 \cup ... \cup T_n) + \Pr(T_2 | T_3 \cup ... \cup T_n)$$

We can define the probability for classes which have any times **and** and **or** by induction. For complex instance, the probability of a gene G which appears in tissue T_1, T_2 and T_3, in tissue T_4 or in tissue T_5 is:

$$\begin{aligned}
&\Pr((T_1 \cap T_2 \cap T_3) \cup T_4 \cup T_5 | T_6 \cup ... \cup T_n) \\
&= \Pr(T_1 | T_6 \cup ... \cup T_n) \times \Pr(T_2 | T_6 \cup ... \cup T_n) \times \Pr(T_3 | T_6 \cup ... \cup T_n) \\
&\qquad + \Pr(T_4 | T_6 \cup ... \cup T_n) + \Pr(T_5 | T_6 \cup ... \cup T_n)
\end{aligned}$$

3 Application to the Design of Micro Array

Bodymap project collects the digital expression profiles. It dissociates human and mouse organ, extracts mRNA, make libraries which include only 3'-end, and registers them. Now we get 18,000 human genes out of 60 organs and 16,000 mouse genes out of 36 organs.

As the way of observing the expression profile, micro arrays are getting popular devices. Although micro arrays are developed as an cost-efficient method to get expressions of numerous genes at once, in practice we further want to reduce the number of cDNAs putting on a micro array. Using our statistical method to Bodymap digital expression profiles, we can list candidates of genes that would be high related with a specific collected tissues ordered by significance of occurrences.

References

[1] Audic, S. and Claverie, J.-M., The significance of digital gene expression profiles, *Genome Res.*, 7:986–995, 1997.

[2] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, *Nature Genet.*, 2(3):173–179, 1992.

[3] Hishiki, T., Kawamoto, S., Morishita, S. and Okubo, K., BodyMap: a human and mouse gene expression database, *Nucleic Acids Res.*, 28(1):136–138, 2000.