

# A Database System for cDNA Expression Profile Using ESTs of Oligo-Capping Clones and UniGene

Naoyuki Harada<sup>1</sup>      Katsuhiko Murakami<sup>2</sup>      Tetsuo Nishikawa<sup>1</sup>  
n-harada@crl.hitachi.co.jp      katsu@ls.hitachi.co.jp      nisikawa@crl.hitachi.co.jp  
Tomoyasu Sugiyama<sup>3</sup>      Toshio Ota<sup>3</sup>      Ryotaro Irie<sup>3</sup>  
sugiyama@hri.co.jp      ota@hri.co.jp      irie@hri.co.jp  
Keiichi Nagai<sup>3</sup>      Takao Isogai<sup>3</sup>  
k-nagai@hri.co.jp      isogai@hri.co.jp

<sup>1</sup> Central Research Laboratory, Hitachi, Ltd., 1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

<sup>2</sup> Life Science Group, Hitachi, Ltd., 1-3-1 Minamidai, Kawagoe, Saitama 350-1165, Japan

<sup>3</sup> Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba 292-0812, Japan

**Keywords:** cDNA, EST, gene expression profile, tissue specificity, database

## 1 Introduction

With progress of the human genome project, the whole sequences of the human genome will be determined completely in a few years. Functional analysis of genes of human genome is now being accelerated. Full-length cDNA clones are now being collected because they can be used as a starting material for functional analyses of genes. The oligo-capping cDNA library developed by Maruyama and Sugano is an effective source of full-length cDNA clones [1]. The aim of this study is to construct an integrated database of full-length cDNA sequences obtained from oligo-capping cDNA library for the functional analysis of genes. For this purpose we introduced a new annotation method using expression profile information obtained from cDNA sequence databases. The database system developed here has a function to retrieve the tissue specific genes. We also proposed a new method that can statistically compare the frequency distributions of gene expression over tissues. Finally the validity of this method was tested using known tissue specific genes. This study is a part of a project to determine the full-length cDNA clones and construct a cDNA database system financed by New Energy and Industrial Technology Developmental Organization (NEDO).

## 2 Database

The cDNA database system consists of the ESTs of cDNA sequences and cDNA library information from which the sequences were extracted. In building the database, we used clustered ESTs of cDNA provided from Helix Research Institute and also clustered ESTs downloaded from UniGene web site [2]. Each cDNA sequence in each cluster has library information. Based on the clusters and the library information, an expression matrix was constructed whose row and column denote each EST cluster and each cDNA library, respectively. The element of the matrix is the relative frequency of sequences that belong to the  $i$ th cluster and were extracted from the  $j$ th library ( $i = 1, \dots, I, j = 1, \dots, J$ ). Because there are great differences among numbers of sequences extracted from each library, the element of the matrix was normalized by the sum of numbers of sequences extracted from  $j$ th library. The system has a function to retrieve tissue specific genes. Given the tissue names (or library names) and expression frequency of genes as a query (Fig. 1), the system output clusters that satisfy the query conditions (Fig. 2). Using this function, we can retrieve the interesting tissue specific genes.

| 1st Tissue Library | 2nd Tissue Library                        | Clone | Select                   |
|--------------------|---|-------|--------------------------|
|                    | Kidney-Kidney, Fetus                      | 1043  | <input type="checkbox"/> |
|                    | Kidney-Kidney, Infant                     | 44    | <input type="checkbox"/> |
|                    | Kidney-Kidney, tumor tissue               | 20448 | <input type="checkbox"/> |
|                    | Larynx                                    | 753   | <input type="checkbox"/> |
|                    | Larynx-Larynx, tumor tissue               | 753   | <input type="checkbox"/> |
| Liver              | Liver-Liver, Adult                        | 8224  | <input type="checkbox"/> |
|                    | Liver-Liver, Fetus                        | 5068  | <input type="checkbox"/> |
|                    | Liver-Liver, Infant                       | 104   | <input type="checkbox"/> |
|                    | Liver-Liver, Microdissected normal tissue | 848   | <input type="checkbox"/> |
|                    | Liver-Liver, Microdissected tumor tissue  | 191   | <input type="checkbox"/> |
|                    | Liver-Liver, tumor tissue                 | 235   | <input type="checkbox"/> |
| Lung               | Lung-Lung, Adult                          | 68    | <input type="checkbox"/> |
|                    | Lung-Lung, Fetus                          | 50848 | <input type="checkbox"/> |
|                    | Lung-Lung, Infant                         | 10071 | <input type="checkbox"/> |
|                    | Lung-Lung, tumor tissue                   | 485   | <input type="checkbox"/> |
|                    | Lung-Lung, tumor tissue                   | 17491 | <input type="checkbox"/> |
|                    | Lung-Lung, tumor tissue                   | 9252  | <input type="checkbox"/> |

Figure 1: cDNA library list.

| 1st Tissue   | 2nd Tissue                                | N.E.rate(%) | EST |
|--------------|---|-------------|-----|
| Liver        | Liver-Liver, Adult                        | 51.71       | 163 |
|              | Liver-Liver, Infant                       | 19.36       | 7   |
| Gall bladder | Gall bladder-Gall bladder                 | 21.35       | 22  |
| Pool         | Pool-Pool, liver+spleen, Infant           | 0.42        | 20  |
|              | Pool-Pool, melanocyte+heart+uterus, Adult | 0.03        | 1   |
| Spleen       | Spleen-Spleen, Fetus                      | 5.72        | 13  |
| Testis       | Testis-Testis                             | 0.74        | 4   |

Figure 2: Expression profile of retrieved gene.

### 3 Method

To compare the gene expression frequency distributions over libraries, and to evaluate the tissue specificity and/or the similarity of the distributions, some methods have been proposed [3, 4]. For more accurate retrieval of tissue specific genes, we propose a statistical measure of tissue specificity of genes. The measure is the following  $\chi^2$ . To compare distributions of  $I_1$ th and  $I_2$ th cluster a  $\chi^2$  statistic is calculated. The similarity of two distributions is represented by  $1 - \text{tail probability } P$  of  $\chi^2$ . Suppose that  $i$  denotes a cluster in the row and  $j$  denotes a library in the column defined in the expression matrix, where  $i = I_1, I_2$  ( $1 \leq I_1, I_2 \leq I$ ) and  $j = 1, \dots, J$ . And  $N_{ij}$  is the  $(i, j)$  element of the expression matrix, then,

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}} \quad \left( n_{ij} = \frac{N_i \cdot N_j}{N}, N_i = \sum_j N_{ij}, N_j = \sum_i N_{ij}, i \in \{I_1, I_2\} \right)$$

$$P(\chi^2 | v) = \int_{\chi^2}^{\infty} e^{-t} t^{\frac{v}{2}-1} dt \quad (v = IJ - I - J + 1)$$

To obtain measure of specificity, a two-by-two frequency table, whose row consists of  $i_1$ , and  $i_2$  EST and column consists of a  $j_1$  library and the mean of the rest of libraries, is made for each  $j_1$  ( $1 \leq j_1 \leq J$ ). Then distributions over  $j_1$  and the mean of the rest of the libraries of  $i_1$  and  $i_2$  EST is compared by  $\chi^2$  statistic similarity.

### 4 Result

We tested validity of our method using known tissue specific genes. We retrieved twenty-nine liver specific genes and sixteen house keeping genes from Eukaryotic Promoter Database (EPD) as a data set [5]. As the result of comparison of the distributions, P value at liver between liver specific genes and house keeping genes are very low. This result showed that our method could measure the tissue specificity of genes.

### Acknowledgment

This work was supported by a Grant from NEDO Project of Ministry of Industrial and Technology of Japan.

### References

- [1] Maruyama, K. and Sugano, S., Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, 138:171–174, 1994.
- [2] <http://www.ncbi.nlm.nih.gov/UniGene/>
- [3] Claverie, J.M., Computational methods for the identification of differential and coordinated gene expression, *Human Molecular Genetics*, 8:1821–1832, 1999.
- [4] Zhang, M.Q., Large-scale gene expression data analysis: a new challenge to computational biologists, *Genome Research*, 9:681–688, 1999.
- [5] Perier, R.C., Praz, V., Junier, T., Bonnar, C., and Bucher, P., The Eukaryotic Promoter Database (EPD), *Nucleic Acids Res.*, 28:302–303, 2000.