Applications of high-throughput identification of tissue expression profiles and specificity

Introduction

Organisms as mammals do not express every single gene encoded by genome in each tissue, although the various cell types of the organism express particular subsets of the genes in the genome. Cell types are further organized into tissues, and tissues constitute the organs that carry out various physiological functions. The detailed mechanisms of gene products underlying the functioning of this complex organization are today largely unknown. Several methods, including SAGE (Vol89) and microarray technology (Sho01) can be applied to the study of differential gene expression in the various cell types, in different tissues. We recently developed TissueInfo, a high-throughput method to identify the tissue expression profiles of the genes in an organism’s genome, as well as the tissue specificity of a query sequence (Bog93). The method carefully organized the data publicly available in dbEST (Bog03) and is purely computational. With 80% coverage of the benchmark collected, TissueInfo achieved an accuracy of 98% of the tissue specificity of a gene is predicted and 89% when its expression in a given tissue is predicted. These results make possible the application of TissueInfo to the complete sequences available in the public draft of the human genome.

Here, we present applications of TissueInfo to genome-wide analysis of tissue expression, gene discovery, construction of tissue targeted microarray slides. Other applications include the assembly of training sets for the ab initio prediction of tissue expression and specificity (promoter analysis).

Materials and Methods

Transcript sequences from the human genome project were obtained from the NCBI (ftp://ncbi.nlm.nih.gov/ genome/ H.sapiens-est). A set of transcript sequences containing 25,612 transscripts in two sets: reference sequences (accession number starting with NM_, 12,685 sequences) which have been manually curated by NCBI staff, and transcript sequences computationally derived from the public assembly of the human genome (accession numbers starting with XM_, 12,927 sequences). As many transcripts overlap among these two datasets, we did the analysis with reference sequences only. Transcript sequences were masked to remove human repeats [Jur00] and used to search the human EST (dbEST) (Timex). The resulting data were filtered with tmatchal as described in the TissueInfo annotation of dbEST (Sho01) with minerror=100 and max-error=0.05. We extracted a list of 104 human tissues from the TissueInfo annotation of dbEST. This list covers most tissues and cells represented in dbEST and adds anatomical groupings of tissues and organs, such as head (contains brain, eye, ear, etc.) or brain (contains hypothalamus, cortex, hippocampus, etc.). We used tiquery to get the tissue information associated with each transcript.

Results and Evaluation

Three figures are grouped under point 1. The histogram on the left shows the number of genes identified expressed in a given tissue. The scatter plot titled “Expression” shows that there is a correlation between the number of genes predicted to be expressed in a given tissue and the number of ESTs present in dbEST, annotated as being prepared from this tissue. On the contrary, the same scatter plot obtained for genes specific to a given tissue shows no discernible correlation. Two tissues lie far below the median line: liver and spleen, suggesting that the estimate of the number of genes expressed in these two tissues will only slightly be improved as ESTs are sequenced from these tissues.

Point 2 shows an evaluation of our results on a test set of 113 genes. Genes were selected to be included in the test set if identified as expressed in all the Tissues of a cluster from the following list: kidney and spleen and liver (24 genes) (placenta and testis and spleen and liver) (1), [skin and heart] and uterus (11) [brain and kidney] (10), [brain and spleen and liver] (7) [lung and liver] (7) [large intestine and brain] (3) [placenta and brain and liver] (2) [placenta and spleen and liver] (6) [brain and spleen and liver] (5) [brain and bone marrow] (5) [lung and gut] (6) [nervous system and brain] (12). Each gene is tested for expression in each tissue belonging to a cluster and the literature is searched to check the identification. In 50% of the cases, the verification is not possible because the information is not available in the literature. In 77% of the cases where verification is possible, TissueInfo was accurate in identifying that a gene was expressed in a given tissue. These results are in good agreement with our previous evaluation of TissueInfo (Sho01).

Gene discovery

TissueInfo in gene discovery as illustrated on point 3. Genes identified to be specific to retina (9) and ear (2) have been shown in greater detail. The 9 genes identified in the tables as specific of retina are verified in the literature. The two genes identified as specific to ear are dentin and a dentin homolog. Dentin is a protein in teeth, but no EST library have been made from teeth. What TissueInfo identifies is therefore a new organ in dentin matrix acidic phosphoprotein (DMP1). “DSSP, a gene encoding dentin sialophosphoprotein is expressed in two proteins: dentin sialoprotein (DSP) and dentin phosphoprotein (DPP)... Notably, missense mutations in DSP are also associated with progressive hearing loss)” [Pat01].

Reference

Not similar to dentin matrix acidic phosphoprotein

Novel statistical features

The histogram shown in point 4 plots the distribution of hits in dbEST for all the genes of the analysis set (genes with less than 3 hits are not shown). The distribution of hits shows two major peaks: one below 10 hits and a second centered on 80 hits. We verified that this feature is conserved in genes specific to a given tissue and noticed that most genes which are specific to a tissue and have more than 80 hits in dbEST are hormones or secreted proteins (6 out of 7 known function). This new type of results is now being analyzed for the possible physiological basis of the b4-modulation.

Construction of custom microarray clone sets

We have shown that the inference produced by application of TissueInfo to genes for which data are not directly available in the literature should be of the same accuracy as those for genes described more completely. In addition, the method is scalable and therefore can be used to complement the data available in the literature and serve as a valuable resource for the selection of clones to build custom microarrays (e.g. liver or brain arrays).