

## Poster I-20

### **Community Bone Microarray Data Annotation System Using Visual Data Analysis Pipeline Framework**

**Shin, Dong-Guk<sup>1</sup>, Gilman, Chris<sup>1</sup>, Krueger, Winfried<sup>2</sup>, Wang, Hsin-Wei<sup>1</sup>, Rowe, David<sup>2</sup>**

**<sup>1</sup>Computer Science and Engineering Department, University of Connecticut, Storrs, CT, USA;**

**<sup>2</sup>Genetics/Developmental Biology Department, University of Connecticut Health Center, Farmington, CT, USA**

Key required features of a community annotation system include data review and data entry/edit methods. Scientists should be able to review the published experiential data using the data review feature and enter their interpretation of the experimental results using the data entry/edit module. There is another essential feature of a community annotation system that is not well understood. That is, during the annotation process, scientists would need to associate the experimental outcome with other related data (e.g., homologous sequence search, pathway look up, PubMed search, etc.). In the conventional environment, such analyses are to be done off-line by supporting bioinformatics staff, because conducting the analysis tasks is too cumbersome for the annotators. In this scenario, data analyses and data annotations are two disparate tasks. We believe that such disconnect makes the annotation process quite inefficient. In a larger scale annotation effort (e.g., a community annotation system), this disparity is even greater, and the inefficiency becomes even more grave.

We propose a novel framework with which scientists can conduct analysis tasks by themselves. In this framework bioinformatics supporting staff publish ready-made analysis pipelines visually. The visual pipelines are easy to understand and greatly help annotators conduct the analysis process by themselves. The visually depicted data analysis pipelines also help annotators review and examine the progress of the analysis tasks in real-time.

Within our current bone microarray annotation system, the annotating scientists can perform the following tasks. (i) Finding Gene Name Alias - During the discovery of a gene, different names were usually assigned to the same gene. Although most of the genes are having a unique “official gene name”, some may be best known by its “alias names”. This process helps the annotator recognize the clone not just by the official gene name, but also any alias that has ever been used to represent the gene. (ii) Finding UniGene History – NCBI’s UniGene is a very useful knowledge base for identifying possible genes. To annotate EST clones, one first needs to associate them with UniGene. By doing this the annotator can obtain more information and better characterize the clone. (iii) Binding Site Analysis - Getting the transcription binding site will help the annotator find the possible upstream and downstream genes in a pathway. (iv) Finding Pathway Partners - Getting the possible pathway partners greatly helps annotators understand group behaviors of genes that may work together.

In summary, we argue that the use of a visual data analysis pipeline can make do-it-yourself data analyses possible for the annotator and in turn significantly increase the speed and quality of annotation.

*This work was supported in part by NIH/NIGMS grant P20 GM 65764-02.*