CANVAS: A Clone Annotation and Validation System

Zhong Li Zhong_Li@sbphrd.com
Junping Jing Junping_2_Jing@sbphrd.com
Jeffrey S. Aaronson Jeffrey_S_Aaronson@sbphrd.com

SmithKline Beecham Pharmaceuticals, 709 Swedeland Rd., King of Prussia, PA 19406

Keywords: microarray, clone, annotation

1 Introduction

When analyzing large-scale gene expression data, it is critical to understand the biological function of genes on the DNA array in order to interpret expression abundance results. However, arrays generally are composed of gene fragments, or clones [Iyer *et al.* 1999], rather than complete genes, and a gene is often redundantly represented within an array. Furthermore, arrayed clones are frequently annotated incorrectly in the public sequence databases [Aaronson *et al.* 1996], and often, are not derived from known genes. Therefore, it becomes necessary to develop and apply computational and informatics techniques that permit gene-centric analyses of microarray clones, and that place these clones into their biological context. We have developed CANVAS (a Clone Annotation and Validation System) in order to address these needs for a microarray grid of human clones.

2 Method and Results

Sequences from human microarray grid clones are regularly searched against a series of sequence databases using BLAST [Altschul *et al.* 1990]. Sequence databases included in CANVAS are: human transcript sequences, mouse transcript sequences, rat transcript sequences, human protein sequences, mouse protein sequences, rat protein sequences, non-human protein sequences, grid clone sequences, and human EST sequences. Computational results are stored in a local relational database. The CANVAS Web interface allows dynamic querying over these results, and returns a detailed report for each clone. Rich interaction between CANVAS and HASTE, a gene expression profile browser, provides access to biological and expression data simultaneously. Queries with both gene-like entities (e.g. a list of mRNA GenBank accessions or UniGene clusters [Boguski and Schuler 1995]) and gene fragment entities (e.g. a list of EST GenBank accessions or clone ids) are permitted. Gene-like queries are embedded within gene fragment query result pages, suggesting other grid clones putatively belonging to the same gene.

CANVAS reports are modular, each component details results from a distinct analysis. Current modules: (1) Report whether internal reads from a clone can be sequence confirmed against public domain reads (ESTs) from the same clone; (2) Provide best available description line annotation based on sequence similarity; (3) Indicate clones with identical sequence both within the grid and within a comprehensive EST database; (4) Suggest homology or identity relationships to known protein and transcript sequences from model organisms; and (5) Place a clone on an integrated physical and genetic map. Hot-links to internal and external Web resources dedicated to functional genomics, disease linkage analysis and gene clustering are also provided.

References

[1] Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S. and Elliston, K.O., Toward the development of a gene index to the human genome, *Genome Res.*, 6(9):829–845, 1996.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215(3):403–410, 1990.

[3] Boguski, M.S. and Schuler, G.D., ESTablishing a human transcript map, *Nature Genet.*, 10(4):369–371, 1995.

[4] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Jr, J.H., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O., The transcriptional program in the response of human fibroblasts to serum, *Science*, 283(5398):83–87, 1999.