Pacific Symposium on Biocomputing 13:513-514(2008)

TILING MICROARRAY DATA ANALYSIS METHODS AND ALGORITHMS

SRINKA GHOSH AND ANTONIO PICCOLBONI
Affymetrix, Inc.
6550 Vallejo st.
Emeryville, CA 94608

The complete sequencing of the human genome and several other genomes for model organisms and other scientifically or technologically important species has opened what has been dubbed the *post-genomic era*. Notwithstanding the continuing and fruitful sequencing projects, this phase has been marked by a strong emphasis on genome function. The promise that the sequence, once revealed, would pave the way to understanding a variety of other aspects of biology has not been fully realized. For instance, the effort to experimentally characterize the structure of proteins is more vigorous then ever and conformation prediction from sequence information alone, despite progress, remains a challenge. Even coming up with a complete gene list for a newly sequenced genome is still a challenge and there is evidence that the transcribed fraction of the genome has been underestimated. Large collaborative efforts, such as the ENCODE project, have been launched to throw an array of experimental technologies at the problem of functional characterization of the genome – including but definitely not limited to more sequencing and in depth comparative genomics. One such technology is the *tiling microarray* (TM). A variation on the now widespread DNA microarray, the TM contains probes that correspond to regularly spaced position on a target genome, irrespective of their annotation as transcripts, promoters or any other functional determination. Therefore, they are made possible by genome sequencing efforts and complement them as high throughput technologies for the characterization of a variety of functional aspects of the genome. In combination with diverse assays, they have been applied to tasks such as transcript mapping and copy number variation and DNA replication analysis. In particular, the combination of TMs with chromatin immunoprecipitation techniques has enabled the high throughput study of protein-DNA binding and chromatin

Pacific Symposium on Biocomputing 13:513-514(2008)

state. With an increasing number of TM-based datasets available to the scientific community, there is a considerable need for improved algorithms and software for their analysis and processing, and this session sought to provide a forum for investigators in the field to present and discuss the most recent advances.

We accepted three papers for this session. In the first, Kuan, Chun and Keles report on some progress in the analysis of chromatin immunoprecipitation TM data. They observe that a correlation structure exists in this type of data and that this aspect has not received enough attention in the literature. They formulate a model that takes this correlation into account and develop a fast detection algorithm based on this model. They support the usefulness of their approach with simulations and a case study and finally provide an open source implementation. In the second, Zeller, Henz, Laubinger, Weigel and Rätsch focus their attention on the application of TM to the characterization of transcription. They offer two related but distinct contributions: one is a normalization method that reduces within-transcript variability, while enhancing the signal separation between exonic and intronic regions; the second is a segmentation method that extends previous work on the unspliced transcript identification problem to the more challenging spliced case. Finally, Danford, Rolfe and Gifford turn the attention away from data analysis to data processing, storage and retrieval. They present a database design for TM data that can handle the results of a variety of experiments and processing methods, can manage multiple species and genome releases and provides convenient graphical presentation for the data, all built on top of a modular architecture amenable to customizations and extensions. They also present a system to formulate and record relationship between different chromatin immunoprecipitation events and provide a reference implementation.