Breast Cancer Genomics Informatics

http://eh3.uc.edu/BreastCancer

Kaustubh Shinde[1], Xiangdong Liu[1,2], Maureen Sartor[1], Prachi Kothiyal[2], Mario Medvedovic[1,2]

[1]Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue ML 56, Cincinnati OH 45267, USA

[2]Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, Cincinnati, OH 45229, USA

Introduction

We describe the project to provide bioinformatics support for breast cancer related genomics initiatives and research projects affiliated with the BCERC at University of Cincinnati. The goal of the project is to provide data management solutions, comprehensive analyses and the portal for disseminating data and results of the analyses for genomics data generated locally. In addition, we aim to construct a "Genomics Context" for analyzing and interpreting new breast cancer genomics data which will be useful to researchers at other BCERC centers and wider research community. The "Genomics Context" consists of a web-accessible comprehensive database of public breast cancer-related microarray data and results of analyses performed on these datasets.

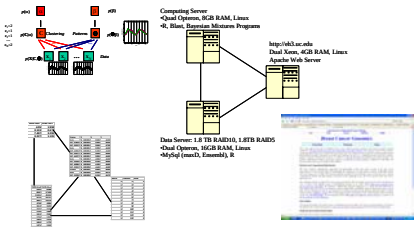
The focus on breast cancer, incorporation of animal and in-vitro model data, and uniqueness of our analytical tools (http://eh3.uc.edu/gimm/csimm) distinguishes this project from other well-known initiatives for organizing cancer genomics data (http://www.oncomine.org).

Computational Infrastructure

➢Three Linux-based servers (Database Server, Computing Server and Web Server)

➢ maxD Microarray Database: open-source MAGE-OM compliant relational database for MySQL (http://bioinf.man.ac.uk/microarray/maxd/)

➢All analytical tasks are performed using open source software: R (http://www.r-project.org/), Bioconductor (http://www.bioconductor.org/) and related packages

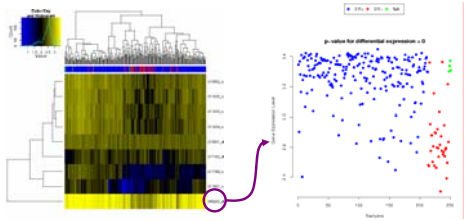


Querying Database

Choosing the dataset

Finding the gene

Selecting gene-related data



Displaying data for all probes for this gene

Displaying data for a selected probe

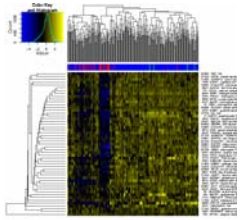
Querying Analysis Results

➢Analysis consists of fitting Bayesian Infinite Mixtures of various complexity (Liu *et al. Bioinformatics* 22:1737-44)

Browsing Clustering Structure using JavaTreeView viewer (Saldanha AJ. Bioinformatics 20: 3246-3248, 2004.).

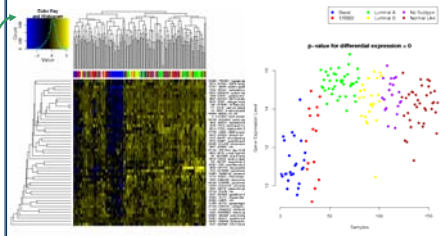
Finding genes "co-expressed" with ESR1 within any dataset based on posterior pairwise probabilities from an infinite mixture model

Clustering genes from the ESR1 neighbourhood within the current dataset using simple correlation-based distances



Combining Different Datasets

Clustering genes from the ESR1 neighbourhood across different datasets and different sample properties



Future Directions

➢Downloading and analyzing more datasets

➢Incorporate locally generated data

➢Development of novel computational procedures for combined analysis of data in different datasets. These efforts are part of the long-term NIH-funded project to develop "Infinite Module Network" models. These models will be capable of simultaneously forming groups of co-expressed genes and identifying and characterizing "contexts" within which such co-expression occurs.

➢Development additional web-based queries (e.g. downloading and uploading gene lists)

➢Development of additional web-based tools for correlating co-expressed genes with functional groupings

Acknowledgements

This work is supported by grants: 1R01HG003749 from NHGRI and U01ES12770 from NIEHS