6.878 Final Report

Dalesh Dharamshi and Iris Xu
(Dated: December 9, 2013)

In the era of personalized medicine it is anticipated that an individual patient's health record will contain multiple molecular expression profiles (genomic, transcriptomic, proteomic, metabolomic, etc) that have been captured across multiple time points in various health states. The combination of this data is expected to provide a much deeper understanding of an individuals healthy and diseased states.

Accurate identification of changes in molecular expression patterns across time and linking them to potential health conditions is at the core of the effectiveness of personalized medicine. Additionally, a patient is likely to have more than one health condition being expressed at the same time.

Unlike a test for differential gene expression in one or more conditions, the clinical setting described above has the additional complexity of analyzing a time course of multiple biological components sampled across disparate time points and health conditions. The challenge is providing a concise description of an individuals state that can be easily analyzed in this clinical setting. We investigate a few aspects of this complexity in order to facilitate generating a clinical recommendation.

While previous work has focused on monitoring expression of specific genes, our goal is to provide an overview of an individuals health state by identifying, separating, and annotating related biological component groups that have similar changes in expression levels, potentially time-shifted indicating a potential causal relationship.

We find that while our method provides visual evidence of effective separation, enrichment analysis shows no significant enrichment between groups of signals.

I. INTRODUCTION

A. Personal Genomics

With the advent of efficient sequencing, the possibility of personalized medicine based on an inidividual's specific genetic composition and molecular phenotype becomes

B. Snyder Dataset

Snyder
Chen, et al. examined how to generate an integrative personal omics profile to examine how biological components change during healthy and diseased states and how to use this information to estimate disease risk and learn about diseased states. While

II. METHODS AND RESULTS

A. Pre-processing

B. Alignment

The goal of our project was to find genes with similar expression patterns, indicating potential biological relevance. In order to first find a measure of similarity, we calculated a distance metric between series using alignment.

Each normalized gene expression series was converted to a character string. We set a threshold change level, initally at 10 percent. If the expression level increased by more than the threshold from the previous timepoint, we convert to an `R` for Rise. If it decreased by more than the threshold, we convert to a `D` for Drop. And if it did not change by more than the threshold level, we convert to `S` for Steady. The first character is always `S`.

For example, the sequence 10, 100, 8, 11, 10 is converted to `S`, `D`, `R`, `S`. Sequences were aligned and scored using Needleman-Wunsch. In order to potentially align signals that have similar but negative

expressions of each other, we attempted to give a smaller mismatch penalty between `R` and `D`. We also give a very large gap penalty for large differences in time. While sequence alignment is polynomial, we aligned on the order of 10^5 sequences, so we modified our inital sequence alignment implementation to include gap penalty lookup as a speedup. Alignment was run on a combination of AWS nodes and the Broad cluster.

C. Clustering

Given the distance matrix generated by alignment, we were able to cluster more similar sequences.

D. Enrichment Analysis

After identifying clusters, we To analyze the efficacy of alignment in clustering, we used Fischer's Exact test for Gene Ontology annotations. Gerald Quon provided us with an `R` script for annotation testing of gene symbols. While, visually, the clusters showed alignment, the p-values indicated no significant enrichment in our clusters.

Additionally, we used `DAVID` Bioinformatics Database web service to find functional groups within each cluster, as a means of describing each cluster, with $kappa = 70$ and $linkage = 0.50$. We were able to generate functional groups for several clusters. The difference between the two

E. Visualization

In order to make expression trend analysis easier from a clinical standpoint, we created a browser-based visualization tool using `d3` that provides normalized/denormalized expression trends by cluster, dendrograms for cluster annotation, and

III. FUTURE GOALS

IV. COMMETARY ON EXPERIENCE

Iris: I learned a lot on how to pull data and to compare results with the ones

V. COMMENTARY ON PEER-REVIEW PROCESS

VI. DIVISION OF LABOR

VII. ACKNOWLEDGEMENTS

Thanks to Gerald Quon, Manolis Kellis, and Maxim Wolf for their help on our project.

---

[1] Chen R, Mias G I, Li-Pook-Than J, Jiang L, et al (2012). Personal Omics profiling reveals dynamic molecular and medical phenotypes. Cell 148, 1293-1307.
[2] Gerald Quon.
[3] Huang DW, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4(1), 44-57.
[4] Huang DW, Sherman BT, Lempicki RA (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1), 1-13.