6.878 Final Report

Iris Xu , Dalesh S Dharamshi
(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 that have been captured across multiple time points in various health states. Accurate identification & representation 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.

We introduce sequence alignment score driven clustering to accurately identify changes in molecular expression over time , this approach can group time-shifted expression signals that may have a causal relationship.We further provide a concept of visualizing these clusters of change in molecular expression across time along with annotations of various health and biologically relevant terms enriched for each cluster.

Our results indicate that using alignment score as a distance to clustering produces clusters that includes signals that may have time shifted due to an upstream or downstream relation.These signals tend to be separated into different clusters while using the standard UPGMA based clustering.

Of the three time course datasets we tried this approach, only two of them appear to show a corresponding enrichment of biological terms associated with the identified clusters. The root cause for the failure for clusters to show significant enrichment in the third dataset is yet to be identified.

I. BACKGROUND

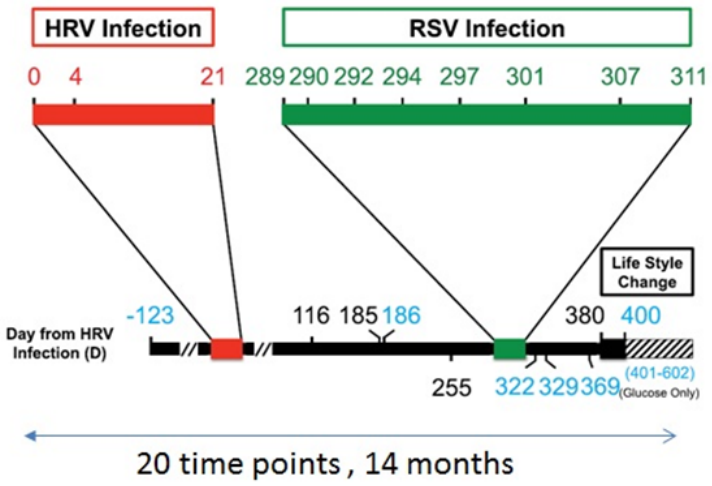

FIG. 1: An individual patients molecular profile captured across time points.

In the era of personalized medicine it is anticipated that an individual patient's health record will contain multiple molecular expression profiles that have been captured across multiple time points in various health states. Accurate identification & representation 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 and grouping molecular expression along the gene or conditions axes to identify changes , our goal is to introduce alignment to potentially improve the accuracy of this grouping by including potentially time-shifted signals that may have a relation. In addition we intend to arrive at an overview of the change in 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.

Note : While our overall project was done as part of two of our classes , in this report we highlight and detail the methods learnt and relevant to this class (Pre-processing , alignment , clustering). We make a brief mention of visualization and annotation for completion.

II. RESULTS AND DISCUSSION

A. Input dataset and Preprocessing

1. GSE33029 : "Integrated analysis of omics profiles"

   We have primarily used this Snyderome dataset which contains various molecular expression captured across 20 time points over 14 months for a single individual.

   The RNASeq dataset posted on GEO is the output of topHat (120GB) and we performed the following preprocessing.

   a. Cufflinks analysis to derive gene expression abundance in FPKMs

   b. Filtering,Quantile normalization,vector normalization of the cufflinks output.

   c. We then created multiple views of this expression data to get a feel for how the data looks.

   In addition to the above we tested our methodology on the following two time course gene expression datasets available on GEO. No additional preprocessing was done to these datasets.

2. GSE19392 "Dynamic responses of primary human bronchial epithelial cells to influenza virus, viral RNA and interferon-beta"

3. GSE675 Time course analysis of response to HCMV infection

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. We wanted similar expression signals to have a smaller distance. In particular, we wanted to capture the fact that changes in expression of one gene may affect the expression of another gene, but in the future. Additionally, we wanted to capture that the upregulation of a gene may cause the downregulation of another gene.

Each processed 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 tried stricter versions of the scoring matrix to allow softened mismatch (between steady and anything else) or no mismatches (Table **??**tab:scoring). We also varied the gap penalty depending on how far apart the time points were. For very far apart signals, we used a larger penalty. 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 similar sequences. Our clustering was based on the alignment scores generated above as distance. One question that needed to

TABLE I: Some sample scoring matrices alignment: least strict (allowing matches between drops and rises in signal for potention downregulation effects), more strict (giving mismatches between rises and drops), most strict (penalizing mismatches). The gap penalty depends on the time offset between two time points, increasing with time differences.

	R	D	S
R	3	-1	-3
D	-1	3	-3
S	-3	-3	1

	R	D	S
R	3	-3	-1
D	-3	3	-1
S	-1	-1	3

	R	D	S
R	3	-3	-3
D	-3	3	-3
S	-3	-3	3

be resolved was how determine the optimal number of clusters.We have tried the following two approaches towards this a. Pairwise clustering until we reach a threshold of the alignment score between two clusters. b. Affinity propagation based clustering to automatically determine the number of clusters.The pairwise heirarchical clustering with an emperical value of threshold based out of experiments gives us fewer clusters than the affinity propogation . However the clusters generated by affinity propogation appear to be more cohesive

We compared our clustering and alignment to the standard UPGMA clustering. We generated a heirarchical clustering output on the same preprocessed dataset with the help of an external tool Cluster 3.0 developed by Michael Eisen at stanford. Shown below are three heatmaps the first two heatmaps are from clusters generated by our implementation of alignment and clustering (affinity prop followed by heirarchical) the third generated by cluster 3.0. In yellow is an example of a cluster that based on the signal appears to belong together but is spread out in multiple clusters if alignment is not considered for clustering.

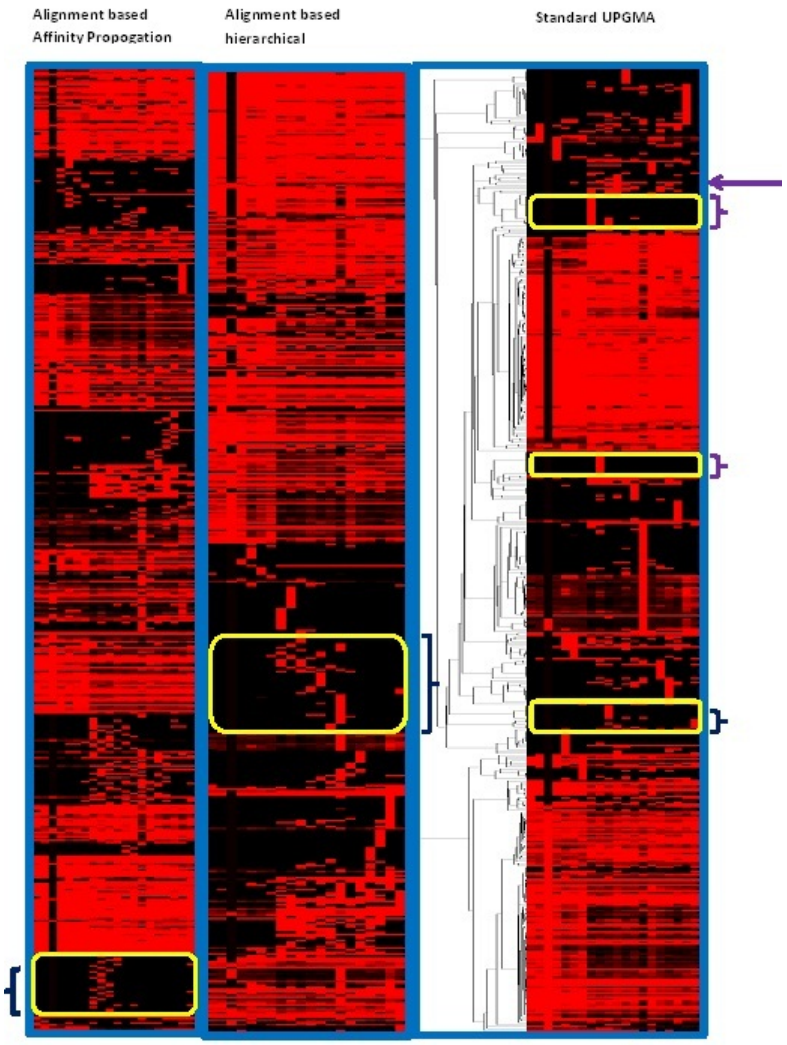

FIG. 2: Heatmap showing the comparison of aligned clustering and UPGMA.

D. Enrichment Analysis

After identifying clusters, we wanted a way to efficiently describe each cluster, as well as to measure the effectiveness of our alignment and clustering algorithm.

To analyze the efficacy of alignment in clustering, we used multiple approaches 1. Fischer's Exact test for Gene Ontology annotations. 2. We ran the gene list generated by our clusters through DAVID functional annotation tool. Gerald Quon provided us with an `R` script and Gene Ontology data for annotation testing of our clusters.

While, visually, the clusters showed alignment for GSE33029 dataset , the p-value matrices indicated no significant enrichment in our clusters after correcting for multiple comparisions with the Benjamini-Hochberg procedure. However, in order to get some quantitative measure of performance, we took sums of the p-value matrices generated from enrichment analysis for our technique for the GSE675 dataset and normalized over the same value for clusters generated by k-means (Fig. **??**fig:enrichment). It shows that, to some degree, our alignment and clustering technique performs better than without alignment.

In order to provide a summarizing description of learned clusters we wanted to provide a word cloud, we initially pulled gene information from the `OMIM` database. However, the text was too general even after multiple round of filtering and the generated word clouds were too non-descriptive.

We then 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 (Fig. **??**fig:enrichment). The `DAVID` clustering provided functional

groups within each cluster, whereas corrected Fischer's Exact test tries to summarize an entire cluster. Since we were able to find significant groups within each cluster, but not for an entire cluster, this may indicate that our clustering is too general or that the clusters are too large.

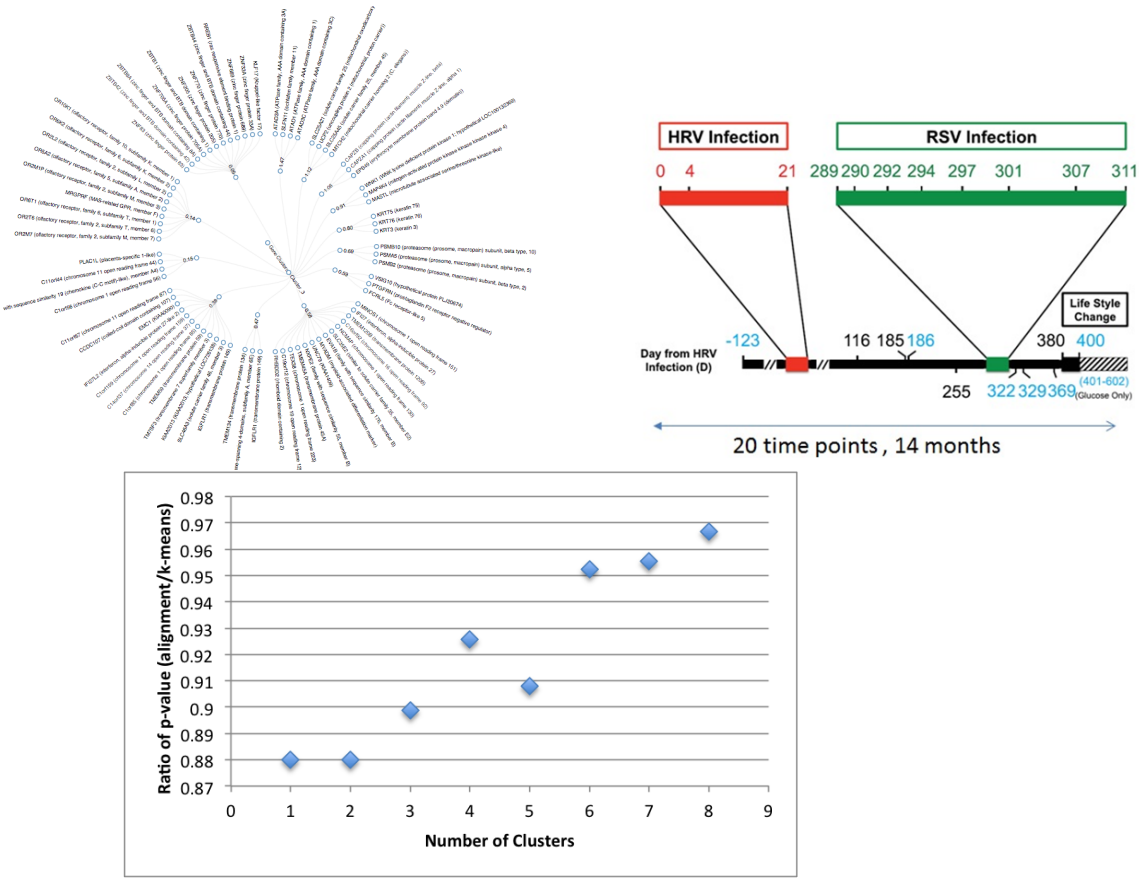

FIG. 3: A) Dendrogram describing a cluster using our visualization tool. B) A word cloud generated from the clusters in the dendrogram. C) To give some performance metric, we took the ratio of the p-values for enrichment analysis for our technique over using plain k-means for the GSE675 dataset.

We have been unable to pinpoint the cause for the lack of enrichment annotations for our clusters on the GSE33029 dataset.We suspected that we are missing a noise correction step for the GSE33029 dataset that may be resulting in no enrichment showing up and reanalyzed with additional normalizations in no vain. We suspect the authors may have not fully specified the steps they took to preprocessing this dataset and we may be missing some correction step.

To check if our approach works we decided to rerun our analysis on two additional and independent timecourse datasets from GEO (GSE19392 GSE675) . With clusters generated from these datasets we do significant enrichment numbers showing up for pathways and other annotation terms. The KEGG pathway results for a few of our clusters are as shown below.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Pathways in cancer	RT		104	3.3	2.6E-5	5.0E-3
	KEGG_PATHWAY	Adherens junction	RT		34	1.1	2.9E-5	2.8E-3
	KEGG_PATHWAY	Viral myocarditis	RT		32	1.0	3.3E-5	2.1E-3
	KEGG_PATHWAY	Focal adhesion	RT		69	2.2	5.3E-5	2.5E-3

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Lysosome	RT		40	1.6	4.0E-6	7.6E-4
	KEGG_PATHWAY	Pathways in cancer	RT		78	3.2	3.6E-4	3.4E-2
	KEGG_PATHWAY	Bladder cancer	RT		17	0.7	5.6E-4	3.5E-2
	KEGG_PATHWAY	Focal adhesion	RT		52	2.1	5.7E-4	2.7E-2

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Ribosome	RT		29	1.2	4.1E-5	7.6E-3
	KEGG_PATHWAY	Spliceosome	RT		32	1.3	3.0E-3	2.5E-1
	KEGG_PATHWAY	Huntington's disease	RT		41	1.7	5.9E-3	3.1E-1
	KEGG_PATHWAY	Oxidative phosphorylation	RT		31	1.3	9.4E-3	3.6E-1

FIG. 4: KEGG Pathway analysis on 3 clusters from GSE19392 dataset using david.

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 word clouds for a text summary of `DAVID` gene groups by cluster (Fig. **??**fig:enrichment).

III. FUTURE GOALS

We would re-do the pre-processing step or attempt to get the pre-processed dataset before reanalyzing our data.

We could integrate this with the SMART Genomics API to make our tool usable for all time series of gene expression.

IV. COMPARISON WITH ORIGINAL PROPOSAL

We managed to complete everything that we proposed, including our alignment technique, several forms of clustering (Affinity Propogation, Hierarchical, K-means), several forms of enrichment analysis (DAVID's function annotations, pulling OMIM text, DAVID's gene grouping, Fischer's Exact test with multiple hypothesis corrections). We also provided a visualization tool as part of our combined class project.

V. COMMETARY ON EXPERIENCE

Iris: I learned a lot on how to pull data and to compare results with the ones in a paper. I also learned a lot about pre-processing data and steps to take when results don't match those of the paper.

Dalesh: Working on the project was an enjoyable experience , a few highlights are as follows a. Opportunity to work in a team and have a mentor to guide us. b. Applying our learning on a couple of real datasets gave additional insights. c. The two interim feedback checkpoints

VI. COMMENTARY ON PEER-REVIEW PROCESS

Iris: I thought that it was good to get feedback, but at the same time, our proposal was missing a lot of detail on purpose because we were not sure at the time exactly on what we were going to do, and a

lot of the feedback was on lack of detail, which was not that helpful. It was nice to see other people's ideas. Dalesh: The peer review process was fun . And i agree with Iris's comment on the feedback we received , i recollect giving ratings on the table without having completely understood what the authors were trying to communicate.

VII. DIVISION OF LABOR

Iris: I did some of the alignment, enrichment analysis, and clustering and helped get jobs running on Broad clusters. I also did the analysis and processing for the GSE675 dataset. Dalesh: Most of the strategy and design was arrived at by brainstorming together , my specific implementation contributions were in data preprocessing on the amazon EC2 instances and clustering , I was responsible for processing and analyzing the GSE19392 dataset , we split the analysis of the GSE33029 dataset . Iris also did the implementations for the user interface.

VIII. 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.
[5] Browne EP, Wing B, Coleman D, Shenk T. Altered cellular mRNA levels in human cytomegalovirus-infected fibroblasts: viral block to the accumulation of antiviral mRNAs. J Virol 2001 Dec;75(24):12319-30. PMID: 11711622