Kevin Chau
A99039092

Network-Based Analysis of Time-Course Differential Gene Expression

Prior to the development of the Network Based Co-Expression Patterns tool (NACEP) by Huang, et al., other differential expression tools did not incorporate temporal gene expression patterns, or if they did, these tools were limited to an independent gene-by-gene approach. NACEP serves to fix this problem by incorporating these co-expression patterns through sophisticated clustering and probabilistic algorithms. This tool was tested and validated on both synthetic datasets and a real control dataset by Ivanova, et al, specifically embryonic stem cell differentiation through retinoic acid.

NACEP works primarily by using co-expression networks to directly compare genes and clustering them together based on their time-course expression patterns. However, instead of strictly assigning genes to single clusters, NACEP generates probabilities of cluster membership, and incorporates these probabilities in differential expression analysis as weights.

NACEP incorporates an infinite-mixture model for clustering the time-course data, in which clusters are assumed to have been generated from a Chinese Restaurant Process. Temporal gene expression patterns are then modeled with a mixed-effects model, such that the expression patterns are represented by the sum of the mean profile of a given cluster (the fixed effect) and biological fluctuations and random technical noise (the random effects). The cluster profiles smooth functions are further modeled as a B-spline whose parameter set is generated through an expectation-maximization process.

In order to find the proper parameters from the data, the NACEP model is written in Bayesian form with a Dirichlet Process prior, which is incorporated into a Gibbs sampling algorithm. This algorithm is run until the parameters converge or a loop-limit is reached. The posterior probabilities generated by the Bayesian model are the probabilities of cluster membership for every gene and every cluster generated by the previous Chinese Restaurant Process.

Differences between conditions for each gene are calculated as weighted distances, such that the weights are the previously-calculated Bayesian posterior probabilities. Statistical significance is calculated through permutation tests, taking into consideration the distribution of the calculated distances; these permutations are used to calculate the false discovery rate, which in turn is used to correct for the multiple hypothesis testing. The overall process is outline in Figure 1.

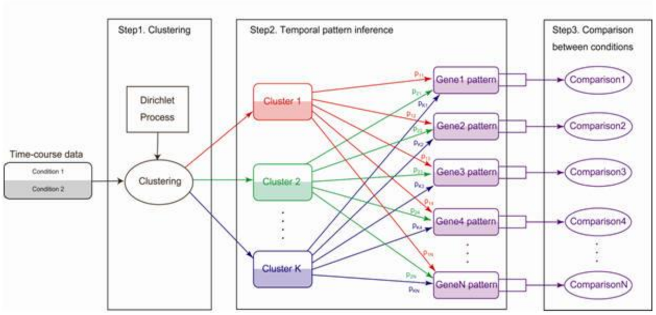

Figure 1. Full NACEP process, beginning from the input time-course data, through clustering, probability calculation, and distance calculations.

Clustering results and temporal pattern comparison results were validated with synthetic datasets and compared to other procedures such as K-means, MClust, and smooth spline clustering (SSC). After one hundred simulated datasets, cluster number detection and gene misclassification results were compared. NACEP outperformed all other applicable tools in both regards, scoring 0% in cluster number prediction error and 0.0733% in average misclassification rate.

Fifty more simulations were conducted with increased variance in gene effect and measurement error to test comparisons of temporal patterns. In this case, NACEP was compared to EDGE, a single-gene based temporal comparison tool. Receiver operating characteristic curves were generated and clearly showed an outperformance of NACEP over EDGE, as shown in Figure 2.

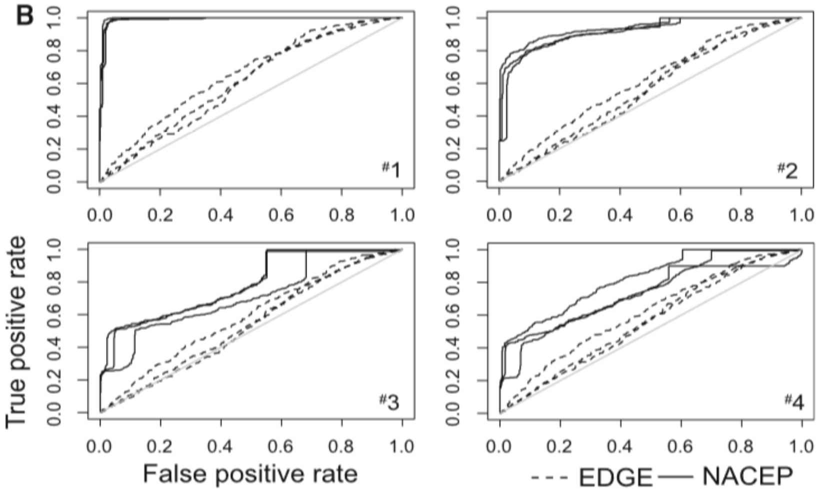

Figure 2. Comparison of ROC curves between NACEP and EDGE.

NACEP was further validated through analysis of a real dataset, provided by Ivanova, et al. In this study, a comparison of spontaneous embryonic stem (ES) cell differentiation versus retinoic acid-induced (RA) differentiation was analyzed such that the authors hypothesized that the temporal expression of the neurogenic regulatory genes that mediate RA-induced ES cell differentiation are different between the two conditions. NACEP ranked Gli3, Zic3 and Shh highly in the RA-induced case, which is consistent with other studies. These results indicate that Shh pathway genes are activated under RA, and this activation has been shown to be associated with neuroplastic growth and brain tumor development.

We then proposed to employ NACEP in our own data analysis. Specifically, we wanted to determine what genes are differentially expressed in prostate cancer. We performed this analysis on publicly available gene expression data provided by the NCBI Gene Expression Omnibus, specifically from the paper “Miz 1, a Novel Target of ING4, Can Drive Prostate Luminal Epithelial Cell Differentiation” by Berger, et al, published in *Prostate* in 2017. The dataset was comprised gene expressions immortalized prostate epithelial stem cells and tumorigenic prostate epithelial stem cells, which overexpressed Erg, Myc, and shPten. These datasets were downloaded as read count matrices with 25,702 genes and six time periods, with three replicates per time period. These count matrices were later converted to TPM expression matrices for use with NACEP.

We next wanted to justify applying NACEP to this data. This was done through a principal components analysis on the samples, whose plot is shown in Figure 3. The first principal component, which captures 97.027% of the variability in the data shows some organization of the data points according to the time periods that they represent, at least in the immortal case. The second principal component also shows the separability of the two conditions.

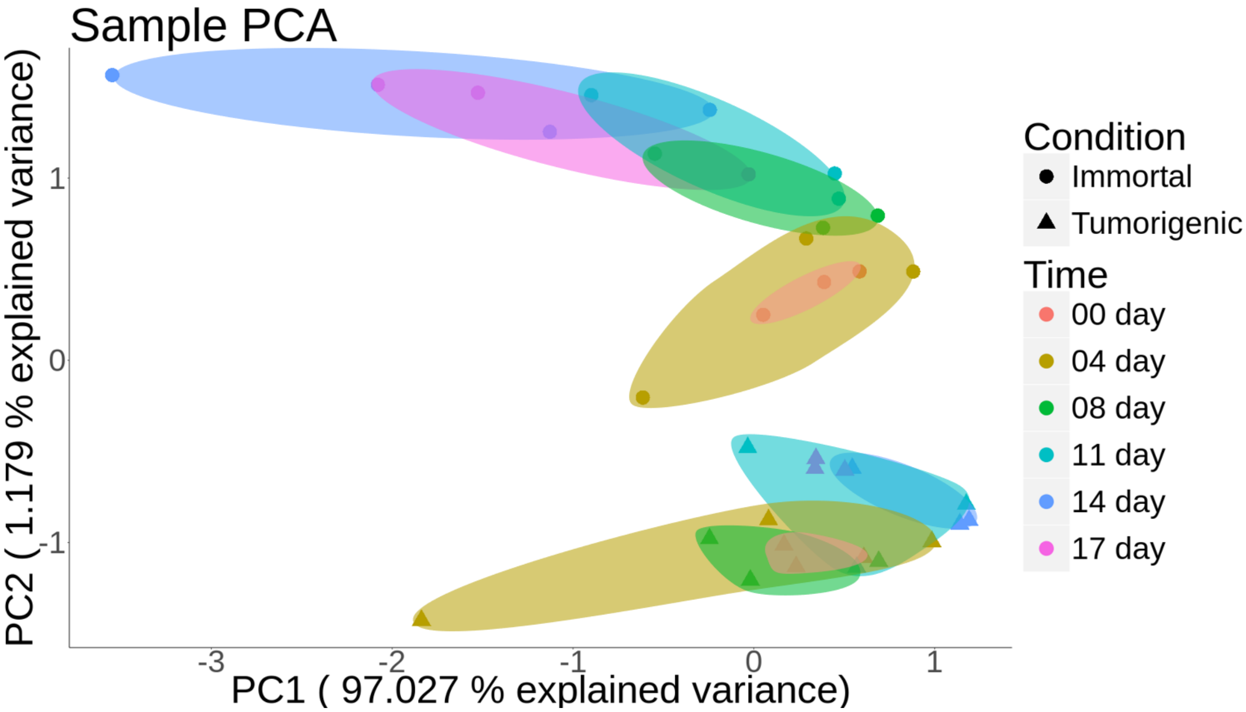

Figure 3. Principal components analysis plot of the first and second principal components.

Next, since NACEP is very computationally intensive, we applied stringent filtering to the TPM expression matrices. We first retained only genes that have 17% of their expression data greater than or equal to 39TPM. Next, we filtered for genes whose log-transformed variance was greater than or equal to 2.4. These values were chosen based on the distribution of the expression matrices themselves. The result was a reduction in gene set size from 25,702 genes to 4,022 genes.

We found several genes with high weighted distances between the two conditions. Among these genes are epithelial cell cytoskeleton genes (KRT14, KRT16), fatty acid biosynthesis genes (SCD), and cell cycle regulation genes (S100A2). Of those high-scoring genes found, many of them have been previously implicated in other prostate cancer studies. However, we found several genes that have not been studied in this context. We, therefore, argue that we have found new genes that could play a role in prostate cancer and that these genes should be further studied in the context of prostate cancer.