1 Hypercubes to Identify Predictive Transcription Factors in a CD4 Yeast Cell-Cycle Genomics Data

We adapted an existing statistical procedure [1], which arranges candidate variables in a k-dimensional hypercube ($k \leq 5$), where the hypercube forms sets of variables for a multi-stage selection process. Although the approach is desirable, it is available currently for traditional models of cross-sectional data. We embedded the hypercube within a dynamic prediction model based on a non-stationary Gaussian linear mixed effects model, in order to accommodate complexity in outcome variable trajectories.

Our proposed method is a practical approach for selecting a handful of truly predictive markers among candidates while accounting for the complex correlations inherent in longitudinal marker data. To show the power of variable selection for the proposed methods, the approach is studied with a simulated data, and is applied to CD4 yeast cell-cycle genomic data, confirming that the proposed method identifies transcription factors that have been highlighted in the literature for their importance as cell cycle transcription factors.

2 Method

Our methodology and details of prediction model are described in our main article [2]. The original article provides additional real data applications with several real cystic-fibrosis (CF) lung function and geomarker datasets and simulations for different data settings. All the computations are implemented in R by using the code provided at GitHub page: https://github.com/emrahgecili/hypercube .

3 CD4 Yeast Cell-Cycle Genomic Data Analysis

We now performed our proposed method on a well-studied yeast cell-circle gene expression data [3-6]. The data were longitudinally collected mRNA gene expression levels in a yeast two cell-cycle period at M/G1-G1-S-G2-M stages. Transcription factors (TFs), which could regulate the gene expression levels during the cell-cycle process, are thus critical to identify. The data we used in this application consists of 297 genes expression levels over 4 time points at G1 stage and $p = 96$ TFs, where TFs are log-transformed.

The proposed method was applied to this data. In our analysis, we first arrange 96 TFs to a $5 \times 5 \times 5$ cube to include all TFs with some 0 entries. Then 43 variables were selected under the choice of keeping predictors with two lowest p-values. Then 43 variables were arranged to form 7×7 two-dimensional square, and with only keeping the variables that are significant at 0.05 level, 14 TFs were identified (CBF1, CIN5, FKH2, GAT3, MBP1, MCM1, NDD1, PUT3, RGM1, RLM1, STB1, STP1, SWI6, YAP5). We tried to fit model (**??**) that includes all 14 selected TFs but the this model failed due to multicollinearity problem. Although the subjects in this study do not have long sequences of repeated measurements, the proposed methods were able to identify important TFs that have already been verified by some biological experiments using genome-wide binding techniques. For example, MBP1 is a crucial transcription factor involved in cell cycle progression from G1 to S stage; NDD1

regulate G2/M genes through binding to their promotes; function of FKH2 is activation of its M stage-specific target genes and it is a cell cycle activator for genes in GFKH2 during the G2 stage; STB1 encodes a protein that contributes to the regulation of SBF and MBF target genes; expression is cell cycle-regulated in stages G1 and S. Our analysis resulted in more discoveries and all of these additional TFs have been reported as key cell cycle TFs in different stages. We refer to studies [3-6] for additional context on this genomic data and TFs.

As aforementioned, our approach corroborated TFs that have been identified previously in the literature for the genomic data example. Note that, the choices of significance level and the dimension of the initial hypercube are not put forward as definitive; significance tests were used informally as an aid to interpretation and are calibrated to decrease the number of candidate variables.

4 References

[1] Cox DR, Battey HS. Large numbers of explanatory variables, a semi-descriptive analysis. Proc Natl Acad Sci U S A. 2017;114(32):8592-8595. doi:10.1073/pnas.1703764114

[2] Cheng Y, Brokamp C, Rasnick E, Kramer EL, Ryan P, Szczesniak RD, Gecili E. Hypercubes to Identify Place-Based Predictors of Rapid Cystic Fibrosis Lung Disease Progression. XX (2024).

[3] Banerjee N, Zhang MQ. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res. 2003;31(23):7024-7031. doi:10.1093/nar/gkg894

[4] Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci U S A. 2005;102(38):13532-13537. doi:10.1073/pnas.0505874102

[5] Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics. 2007;23(12):1486-1494. doi:10.1093/bioinformatics/btm125

[6] Gecili E, Sivaganesan S, Asar O, Clancy JP, Ziady A, Szczesniak RD. Bayesian regularization for a nonstationary Gaussian linear mixed effects model. Stat Med. 2022;41(4):681-697. doi:10.1002/sim.9279