INTEGRATIVE DATA-DRIVEN MATHEMATICAL MODELS PREDICT NOVEL GENOME-SCALE CORRELATION BETWEEN DNA REPLICATION INITIATION AND RNA TRANSCRIPTION DURING THE CELL CYCLE IN YEAST

Orly Alter(a)*, Gene H. Golub(b), Patrick O. Brown(c) and David Botstein(d)
(a)Department of Biomedical Engineering and Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, Departments of (b)Computer Science and (c)Biochemistry, Stanford University, Stanford, CA 94305, and (d)Lewis-Sigler Institute of Genomics, Princeton University, Princeton, NJ 08544
*orlyal@mail.utexas.edu

INTRODUCTION. Integrative analysis of genome-scale molecular biological signals, such as mRNA expression and proteins' DNA-binding occupancy levels, that correspond to activities of cellular systems, such as DNA replication and RNA transcription, promises to give new insights into mechanisms of regulation. Such integrative analysis requires mathematical tools that are able to represent any number of large-scale datasets in terms of a chosen dataset, which is designated the "basis" set, while reducing the complexity of the data to make them comprehensible. Moreover, these tools should provide data-driven mathematical models for the description of the data, where the variables and operations may represent some biological reality. Recently we showed that singular value decomposition (SVD [1] and generalized SVD (GSVD) [2] provide such data-driven models for genome-scale molecular biological data. In the analyses of yeast cell cycle time courses RNA transcription data [3], the variables of SVD, "eigengenes" and "eigenarrays," and these of GSVD, "genelets" and "arraylets," were shown to correlate with observed genome-scale effects of cell cycle regulators and measured samples of the cell cycle stages that they regulate, respectively. Mathematical classification of genes and arrays according to their expression of these eigengenes and eigenarrays, or genelets and arraylets, outlined the progression of the cell cycle along genes and in time, respectively.

Now we show that pseudoinverse projection [4] provides an integrative data-driven model for genome-scale molecular biological data. We illustrate this model by integrating yeast genome-scale proteins' DNA-binding data of nine cell cycle transcription factors [5], and four DNA replication initiation proteins [6], into the cell cycle RNA transcription time course data, using as basis sets the SVD- and GSVD-determined eigenarrays and arraylets.

METHODS. The proteins' DNA-binding dataset we analyze tabulates the relative DNA-bound protein occupancy levels of the 2,928 ORFs with a P-value <0.1 for at least one data point in any one of the 13 samples: nine yeast cell cycle transcription factors measured by Simon et al. [8] and four yeast replication initiation proteins measured by Wyrick et al. [9].

We reconstruct the proteins' DNA-binding data in the SVD- and GSVD-cell cycle RNA transcription bases using pseudoinverse projections in the intersections of 2,227 and 2,139 ORFs, out of which only 400 and 377 were microarray-classified as cell cycle regulated [3], and 58 and 60 by traditional methods, respectively. These reconstructions least-squares-approximate each binding profile as a linear superposition of the expression patterns of the eigenarrays and arraylets that span the SVD- and GSVD-cell cycle bases, respectively. We then map the reconstructed binding profiles onto the SVD- and GSVD-subspaces, associating with each profile cell cycle phase and amplitude. Independently, we parallel- and antiparallel

associate each binding profile with most likely cell cycle stages, or none thereof, using combinatorics and assuming hypergeometric probability distribution [10] of the 506 and 77 ORFs, that were microarray- and traditionally-classified as cell cycle-regulated, respectively, among all 2,928 ORFs.

RESULTS. With the nine transcription factor samples ordered MBP1, SWI4, SWI6, FKH1, FKH2, NDD1, MCM1, ACE2 and SWI5, following Simon et al., and the ORFs sorted according to their SVD- and GSVD-cell cycle phases, the SVD- and GSVD-reconstructed transcription factors' data approximately fit traveling waves, cosinusoidally varying across the ORFs as well as the nine samples. These traveling waves outline the progression of cell cycle transcription along the genes and in time as it is regulated by the binding of the factors at the promoter regions of the transcribed genes. Simon et al. observed a similar traveling wave in the binding data of these nine factors, across only 213 ORFs with a P-value <0.001 for at least one data point in any one of the nine samples, that were microarray-classified as cell cycle-regulated, sorted according to their cell cycle phases as calculated by Spellman et al. The reconstructed replication initiation proteins' data approximately fit standing waves, cosinusoidally varying across the ORFs and constant across the four samples, antiparallel to the reconstructed profiles of MBP1, SWI4 and SWI6.

The SVD- and GSVD-mapping of the binding profiles onto the cell cycle transcription subspaces are consistent with the probabilistic associations by ORF annotations. For the nine transcription factors, they are also in agreement with the current understanding of the cell cycle program. For example, the mathematical mapping of the profiles of MBP1, SWI4 and SWI6 onto the cell cycle stage G1 corresponds to the biological coordination between the binding of these factors to the promoter regions of ORFs and the subsequent peak in transcription of these ORFs during G1. The profiles of the replication initiation proteins are mapped onto the cell cycle stage, that is the antipodal stage of G1. For the first time it is predicted that the bindings of MCM3, MCM4, MCM7 and ORC1 to DNA regions adjacent to ORFs are correlated with minima or even shutdown in the transcription of these ORFs during G1.

DISCUSSION. Initiation of DNA replication requires binding of ORC and MCM proteins at origins of replications across the yeast genome during G1 [8]. Thus, either one of two mechanisms of regulation may be underlying this novel genome-scale correlation between DNA replication initiation and RNA transcription during the yeast cell cycle: The binding of replication initiation proteins to origins of replication may repress, or even shut down, the transcription of adjacent genes. Or, the transcription of genes may reduce the binding efficiency of adjacent origins. This is the first time that data-driven mathematical models have been used to predict a genome-scale biological principle.

REFERENCES.

1. Alter et al. (2000) Proc. Natl. Acad. Sci. U.S.A. **97**, 10101.
2. Alter et al. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 3351.
3. Spellman et al. (1998) Mol. Biol. Cell. **9**, 3273.
4. Golub and Van Loan (1996) Matrix Computation (Johns Hopkins Univ. Press).
5. Simon et al. (2001) Cell **106**, 697.
6. Wyrick et al. (2001) Science **294**, 2357.
7. Tavazoie et al. (1999) Nat. Genet. **22**, 281.
8. Kelly and Brown (2000) Annu. Rev. Biochem. **69**, 829.