In: M. P. Deutscher, S. Black, P. E. Boehmer, G. D'Urso, T. Fletcher, F. Huijing, A. Marshall, B. Pulverer, B. Renault, J. D. Rosenblatt, J. M. Slingerland and W. J. Whelan, editors, Miami Nature Biotechnology Winter Symposium: Cell Cycle, Chromosomes and Cancer. Miami Beach, FL: University of Miami School of Medicine, vol. 15 (January 31 – February 4, 2004); http://www.med.miami.edu/mnbws/documents/Alter-.pdf.

NOVEL GENOME-SCALE CORRELATION BETWEEN DNA REPLICATION AND RNA TRANSCRIPTION DURING THE CELL CYCLE IN YEAST IS PREDICTED BY DATA-DRIVEN MODELS

Orly Alter [(a)*], Gene H. Golub[(b)], Patrick O. Brown[(c)] and David Botstein[(d)]
Departments of [(a)]Genetics, [(b)]Computer Science and [(c)]Biochemistry, Stanford, CA 94305 and [(d)]Lewis-Sigler Institute of Genomics, Princeton, NJ 08544

*Current address: Department of Biomedical Engineering and Institute for Cellular and Molecular Biology, University of Texas at Austin, TX 78712
orlyal@mail.utexas.edu

INTRODUCTION. Recently we showed that singular value decomposition (SVD) (1) and generalized SVD (GSVD) (2), applied to genome-scale datasets of yeast RNA transcription during the cell cycle (3), provide data-driven models, i.e., mathematical frameworks for the description of the data, where the mathematical variables and operations may represent biological reality. The variables of SVD, "eigengenes" and "eigenarrays," and these of GSVD, "genelets" and "arraylets," appear to represent independent processes and corresponding cellular states (such as observed genome-scale effects of cell cycle regulators and measured samples in which these regulators are overactive), respectively. Mathematical reconstruction of gene and array expression in a subset of eigengenes and eigenarrays, or that of genelets and arraylets, appears to simulate experimental observation of only the process and cellular state, respectively, that these expression patterns represent. Now we incorporate into these models genome-scale datasets of yeast protein binding, including nine cell cycle transcription factors (4), and four DNA replication initiation proteins (5), across 2,928 ORFs.

METHOD. SVD and GSVD applied to yeast cell cycle expression datasets (3) determined the expression patterns of two eigengenes and corresponding eigenarrays (across 4,579 ORFs) (1) and six genelets and corresponding arraylets (across 4,523 ORFs) (2), respectively, that span the SVD- and GSVD-yeast cell cycle transcription subspaces. We map the protein binding dataset (4, 5) onto the SVD- and GSVD-subspaces using pseudoinverse projections (6) in the intersections of 2,139 and 2,227 ORFs, respectively, associating with each protein binding profile cell cycle phase and amplitude. We also parallel- and antiparallel-associate each binding profile with a most probable cell cycle stage (or none thereof) using combinatorics and assuming hypergeometric distribution (7) of the 506 and 77 ORFs, that were microarray- and traditionally-classified as cell cycle-regulated (3), respectively, among all 2,928 ORFs.

RESULTS. The SVD- and GSVD- mapping of the binding profiles onto the cell cycle transcription subspaces are consistent with the probabilistic associations by ORF annotations (see supplemental Fig. 1 and Table 1 online). The correlations of the binding profiles of the nine cell cycle transcription factors with stages of the cell cycle are in agreement with the current understanding of the yeast cell cycle program (4). The genome-scale binding profiles of *ORC1*, *MCM3*, *MCM4*, and *MCM7* are correlated with transcription minima during the cell cycle stage G1.

DISCUSSION. The mapping of the genome-scale binding profiles of the four DNA replication initiation proteins onto a state of transcription minima during G1, predicted by both the SVD and GSVD data-driven models, and supported by the most probable associations by ORF annotations, suggests the following genome-scale correlation between DNA replication and RNA transcription during the yeast cell cycle, that has not been demonstrated before: The binding of ORC and MCM proteins during G1, which is known to be required for initiation of replication at origins of replications across the yeast genome (8), is correlated with a significant reduction in, or maybe even a shut-down of the transcription of those ORFs, to which the ORC and MCM proteins bind. This is the first time that a data-driven mathematical model has been used to make a genome-scale biological prediction.

REFERENCES

1. Alter, O. et al. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101–10106.
2. Alter, O. et al. (2003) *Proc. Natl. Acad. Sci. U.S.A.* 100, 3351–3356.
3. Spellman, P.T., Sherlock, G. et al. (1998) *Mol. Biol. Cell.* 9, 3273–3297.
4. Simon, I. et al. (2001) *Cell* 106, 697–708.
5. Wyrick, J.J., Aparicio, J.G. et al. (2001) *Science* 294, 2357–2360.
6. Golub, G.H., and Van Loan, C.F. (1996) *Matrix Computation* (JHU Press).
7. Tavazoie, S. et al. (1999) *Nat. Genet.* 22, 281–285.
8. Kelly, T.J., and Brown, G.W. (2000) *Annu. Rev. Biochem.* 69, 829–880.