Published in the Proceedings of the 2004 Miami Nature Biotechnology Winter Symposium on the Cell Cycle, Chromosomes and Cancer, edited by Deutscher, M.P. et al. (University of Miami School of Medicine, Miami, 2004), Vol. 15, http://www.med.miami.edu/mnbws/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.

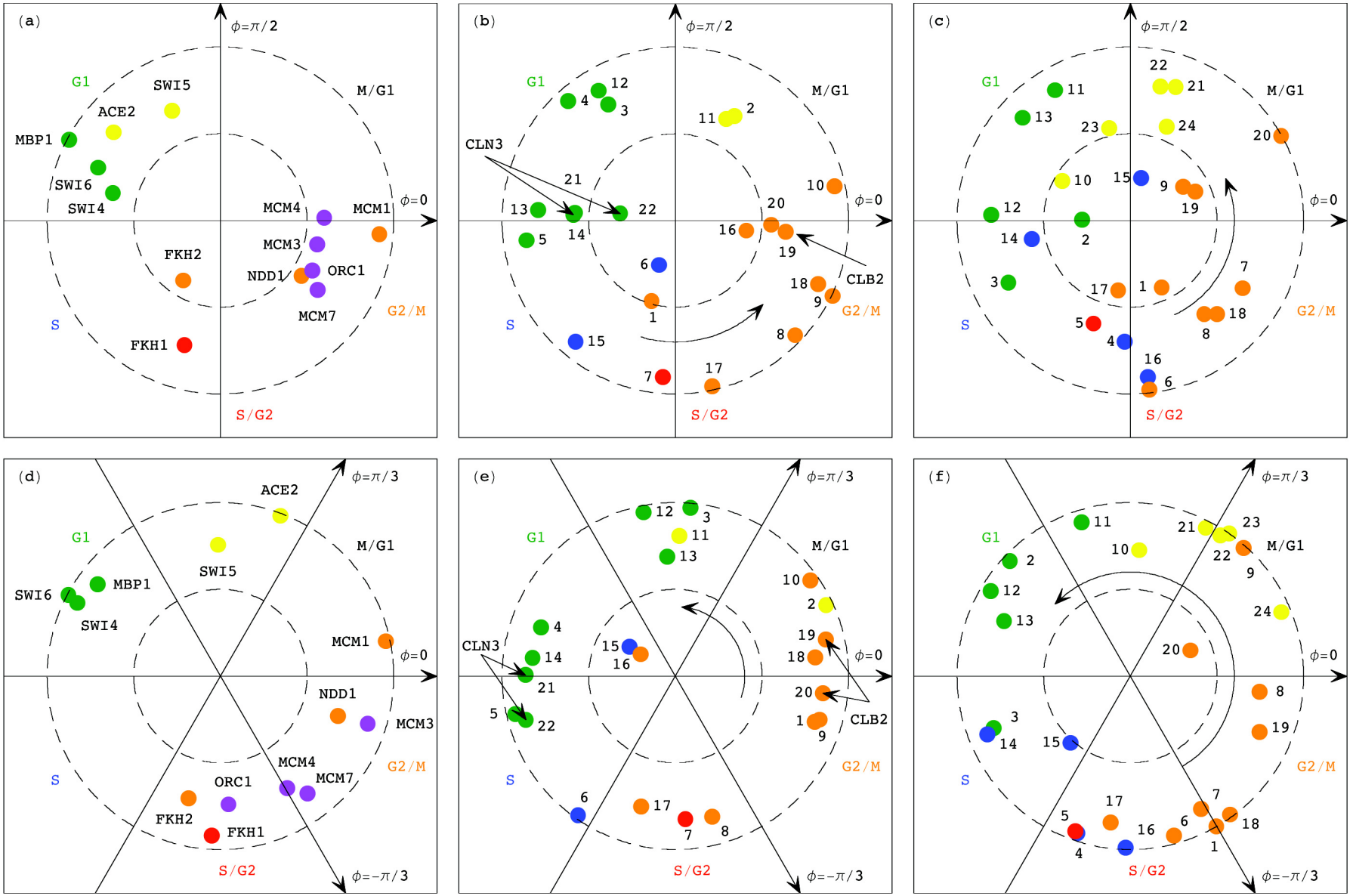
RESULTS. The SVD- and GSVD- mapping of the binding profiles onto the cell cycle transcription subspaces are consistent with the probabilistic associations by ORF annotations (see supplemental Fig. 1 and Table 1 online). The correlations of the binding profiles of the nine cell cycle transcription factors 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.

Fig. 1. Mapping of genome-scale datasets onto SVD- and GSVD-cell cycle transcription subspaces.



Projections of (*a*) the protein binding dataset (4, 5), (*b*) the α factor cell cycle expression time course and overexpression of cell cycle regulators, and (*c*) the *CDC15* cell cycle expression time course (3), onto the two-dimensional SVD cell cycle subspace (1), spanned by the 0-phase eigenarray along the x-axis and the π/2-phase eigenarray along the y-axis, color coded according to the associations by ORF annotations with either one of the five cell cycle stages: M/G1 (yellow), G1 (green), S (blue), S/G2 (red) and G2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall array expression or binding information in this SVD cell cycle subspace.

Projections of (*d*) the protein binding dataset (4, 5), (*e*) the α factor cell cycle expression time course and overexpression of cell cycle regulators, and (*f*) the *CDC15* cell cycle expression time course (3), onto the six-dimensional GSVD cell cycle subspace (2), approximated by the two-dimensional subspace spanned by the 0-phase along the x-axis and the π/2-phase along the y-axis, color coded according to the associations by ORF annotations with either one of the five cell cycle stages: M/G1 (yellow), G1 (green), S (blue), S/G2 (red) and G2/M (orange). The dashed unit and half-unit circles outline 100% and 50% of added-up (rather than cancelled-out) contributions of the six arraylets to the overall projected array expression or binding information. The arrows describe the projections of the –π/3-, 0-, and π/3-phase arraylets onto this approximated subspace.

Table 1. Associations of genome-scale protein binding profiles and expression patterns by microarray- and traditional-ORF annotations.

				Microarray annotation				Traditional annotation			
				Parallel association		Antiparallel association		Parallel association		Antiparallel association	
	Dataset	Array		Most likely	P-value of	Most likely	P-value of	Most likely	P-value of	Most likely	P-value of
a	Binding of cell cycle transcription factors		*MBP1*	G_1	1.6×10^{-14}	None	4.0×10^{-3}	G_1	2.7×10^{-10}	None	9.3×10^{-2}
			SWI4	G_1	1.5×10^{-17}	None	1.2×10^{-1}	G_1	2.7×10^{-7}	None	9.3×10^{-2}
			SWI6	G_1	4.7×10^{-32}	G_2/M	7.3×10^{-2}	G_1	4.8×10^{-19}	G_2/M	4.4×10^{-2}
			FKH1	S/G_2	7.2×10^{-4}	None	3.5×10^{-1}	S/G_2	4.0×10^{-2}	S	3.9×10^{-1}
			FKH2	G_2/M	3.9×10^{-11}	None	8.3×10^{-2}	G_2/M	3.7×10^{-6}	None	2.7×10^{-2}
			NDD1	G_2/M	2.0×10^{-19}	G_1	9.5×10^{-2}	G_2/M	5.0×10^{-9}	M/G_1	3.3×10^{-1}
			MCM1	G_2/M	1.2×10^{-12}	G_1	4.0×10^{-3}	G_2/M	1.6×10^{-7}	G_1	3.3×10^{-2}
			ACE2	M/G_1	1.1×10^{-3}	G_2/M	8.4×10^{-3}	M/G_1	1.1×10^{-1}	S	7.8×10^{-2}
			SWI5	M/G_1	1.3×10^{-15}	G_1	4.5×10^{-5}	M/G_1	6.2×10^{-4}	G_2/M	6.2×10^{-5}
b	Binding of DNA replication initiation proteins		*ORC1*	None	4.0×10^{-3}	G_1	4.3×10^{-13}	None	2.2×10^{-1}	G_1	5.0×10^{-4}
			MCM3	None	4.5×10^{-4}	G_1	7.9×10^{-10}	None	2.7×10^{-2}	G_1	5.0×10^{-4}
			MCM4	None	1.3×10^{-2}	G_1	1.2×10^{-8}	None	4.0×10^{-3}	G_1	2.4×10^{-3}
			MCM7	None	1.3×10^{-2}	G_1	7.9×10^{-10}	None	2.7×10^{-2}	G_1	5.0×10^{-4}
c	α factor cell cycle expression time course	1	0 min	G_2/M	3.2×10^{-6}	G_1	4.9×10^{-27}	M/G_1	4.4×10^{-6}	G_1	7.0×10^{-14}
		2	7 min	M/G_1	5.7×10^{-4}	S	1.3×10^{-6}	M/G_1	1.3×10^{-2}	S	3.4×10^{-6}
		3	14 min	G_1	4.3×10^{-26}	G_2/M	3.2×10^{-6}	G_1	4.2×10^{-7}	S	3.4×10^{-6}
		4	21 min	G_1	3.9×10^{-57}	G_2/M	1.1×10^{-18}	G_1	7.0×10^{-14}	M/G_1	1.7×10^{-7}
		5	28 min	G_1	4.5×10^{-19}	G_2/M	6.2×10^{-16}	G_1	2.0×10^{-11}	M/G_1	4.5×10^{-9}
		6	35 min	S	2.1×10^{-10}	M/G_1	2.1×10^{-20}	S	3.4×10^{-6}	M/G_1	4.4×10^{-6}
		7	42 min	S/G_2	1.2×10^{-11}	M/G_1	4.8×10^{-25}	S	1.0×10^{-2}	M/G_1	1.1×10^{-12}
		8	49 min	G_2/M	6.2×10^{-16}	M/G_1	4.7×10^{-30}	G_2/M	7.6×10^{-3}	M/G_1	1.1×10^{-12}
		9	56 min	G_2/M	3.1×10^{-31}	G_1	6.8×10^{-51}	G_2/M	2.7×10^{-8}	G_1	3.5×10^{-15}
		10	63 min	G_2/M	6.2×10^{-16}	G_1	6.7×10^{-16}	G_2/M	5.7×10^{-4}	G_1	4.2×10^{-8}
		11	70 min	M/G_1	1.6×10^{-21}	S/G_2	4.1×10^{-9}	M/G_1	1.7×10^{-7}	S	3.4×10^{-6}
		12	77 min	G_1	5.1×10^{-61}	S/G_2	1.4×10^{-7}	G_1	2.3×10^{-22}	S/G_2	5.5×10^{-3}
		13	84 min	G_1	5.7×10^{-34}	G_2/M	1.4×10^{-21}	G_1	1.6×10^{-16}	G_2/M	1.1×10^{-6}
		14	91 min	G_1	1.8×10^{-8}	G_2/M	8.4×10^{-9}	S	3.4×10^{-6}	G_2/M	6.6×10^{-2}
		15	98 min	S/G_2	3.3×10^{-3}	M/G_1	2.0×10^{-3}	S	3.4×10^{-6}	M/G_1	1.3×10^{-2}
		16	105 min	G_2/M	4.8×10^{-4}	M/G_1	3.3×10^{-18}	G_2/M	7.6×10^{-3}	M/G_1	8.8×10^{-5}
		17	112 min	G_2/M	3.1×10^{-10}	M/G_1	3.3×10^{-18}	S/G_2	5.5×10^{-2}	G_1	3.8×10^{-6}
		18	119 min	G_2/M	3.3×10^{-14}	G_1	9.6×10^{-21}	G_2/M	3.0×10^{-5}	G_1	4.2×10^{-8}
d	Overexpression of cell cycle regulators	19	*CLB2*	G_2/M	2.1×10^{-67}	G_1	7.3×10^{-26}	G_2/M	7.0×10^{-14}	G_1	9.2×10^{-9}
		20	*CLB2*	G_2/M	1.0×10^{-55}	G_1	1.2×10^{-31}	G_2/M	7.0×10^{-14}	G_1	2.0×10^{-14}
		21	*CLN3*	G_1	1.2×10^{-61}	M/G_1	6.5×10^{-13}	G_1	5.2×10^{-27}	G_2/M	1.0×10^{-4}
		22	*CLN3*	G_1	5.1×10^{-48}	G_2/M	4.5×10^{-18}	G_1	1.1×10^{-15}	G_2/M	1.0×10^{-4}

				Microarray annotation				Traditional annotation			
				Parallel association		Antiparallel association		Parallel association		Antiparallel association	
	Dataset	Array		Most likely	P-value of	Most likely	P-value of	Most likely	P-value of	Most likely	P-value of
e	*CDC15*	1	10 min	M/G_1	1.1×10^{-3}	G_2/M	5.2×10^{-8}	G_2/M	1.4×10^{-2}	S	2.1×10^{-2}
	cell cycle	2	30 min	G_1	1.4×10^{-12}	G_2/M	4.1×10^{-21}	G_1	2.9×10^{-5}	G_2/M	1.4×10^{-3}
	expression	3	50 min	G_1	2.3×10^{-29}	G_2/M	1.5×10^{-30}	G_1	4.9×10^{-14}	G_2/M	1.6×10^{-7}
	time course	4	70 min	S	2.5×10^{-8}	M/G_1	2.8×10^{-19}	S	2.6×10^{-5}	M/G_1	1.6×10^{-5}
		5	80 min	S/G_2	2.2×10^{-9}	M/G_1	9.6×10^{-15}	S	1.0×10^{-3}	M/G_1	7.6×10^{-7}
		6	90 min	G_2/M	3.8×10^{-20}	G_1	3.0×10^{-20}	G_2/M	9.3×10^{-5}	G_1	3.2×10^{-9}
		7	100 min	G_2/M	3.8×10^{-20}	G_1	2.3×10^{-29}	G_2/M	9.3×10^{-5}	G_1	3.2×10^{-9}
		8	110 min	G_2/M	2.0×10^{-29}	G_1	1.9×10^{-27}	G_2/M	3.9×10^{-9}	G_1	2.4×10^{-10}
		9	120 min	G_2/M	9.3×10^{-15}	G_1	6.8×10^{-12}	G_2/M	1.4×10^{-3}	S	2.6×10^{-5}
		10	130 min	M/G_1	4.2×10^{-18}	S	4.5×10^{-3}	M/G_1	7.6×10^{-7}	S	1.0×10^{-3}
		11	140 min	G_1	3.0×10^{-20}	S/G_2	8.6×10^{-8}	G_1	3.7×10^{-6}	G_2/M	1.4×10^{-3}
		12	150 min	G_1	1.3×10^{-18}	G_2/M	4.7×10^{-5}	G_1	3.8×10^{-8}	G_2/M	1.4×10^{-2}
		13	160 min	G_1	6.8×10^{-12}	G_2/M	4.7×10^{-5}	G_1	4.0×10^{-7}	G_2/M	1.4×10^{-3}
		14	170 min	G_1	1.2×10^{-4}	M/G_1	2.7×10^{-4}	S	2.6×10^{-5}	G_2/M	1.4×10^{-2}
		15	180 min	S	1.1×10^{-3}	M/G_1	5.9×10^{-5}	S	1.0×10^{-3}	None	3.5×10^{-1}
		16	190 min	G_2/M	4.7×10^{-4}	G_1	1.0×10^{-14}	S	1.0×10^{-3}	M/G_1	2.7×10^{-3}
		17	200 min	G_2/M	4.6×10^{-10}	G_1	1.5×10^{-5}	G_2/M	1.4×10^{-3}	M/G_1	2.2×10^{-2}
		18	210 min	G_2/M	1.5×10^{-11}	G_1	3.0×10^{-20}	G_2/M	1.4×10^{-3}	G_1	3.2×10^{-9}
		19	220 min	G_2/M	2.3×10^{-17}	G_1	1.5×10^{-5}	G_2/M	1.4×10^{-3}	G_1	8.7×10^{-2}
		20	230 min	G_2/M	1.4×10^{-5}	G_1	5.0×10^{-6}	G_2/M	9.3×10^{-5}	G_1	8.7×10^{-2}
		21	240 min	G_2/M	1.1×10^{-8}	S	4.9×10^{-2}	M/G_1	2.4×10^{-4}	S	2.2×10^{-1}
		22	250 min	M/G_1	1.2×10^{-11}	S/G_2	1.2×10^{-1}	M/G_1	2.8×10^{-8}	S/G_2	1.3×10^{-2}
		23	270 min	M/G_1	1.2×10^{-5}	G_2/M	3.5×10^{-3}	M/G_1	1.6×10^{-5}	G_2/M	1.4×10^{-2}
		24	290 min	M/G_1	5.9×10^{-5}	G_2/M	4.7×10^{-4}	M/G_1	2.4×10^{-4}	G_2/M	1.4×10^{-2}

Most likely parallel- and antiparallel-associations by microarray- and traditional-annotations of yeast ORFs as cell cycle regulated (3) of (*a*) the binding profiles of cell cycle transcription factors (4), (*b*) the binding profiles of DNA replication initiation proteins (5), (*c*) the α factor cell cycle expression time course, (*d*) the overexpression of cell cycle regulators, and (*e*) the *CDC15* cell cycle expression time course (3).