

# The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication

Aviva Presser\*<sup>†</sup>, Michael B. Elowitz<sup>‡</sup>, Manolis Kellis<sup>†§</sup>, and Roy Kishony\*<sup>¶||</sup>

\*School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; <sup>†</sup>Broad Institute, Cambridge, MA 02142; <sup>‡</sup>Division of Biology and Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125; <sup>§</sup>Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and <sup>¶</sup>Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Edited by Leonid Kruglyak, Princeton University, Princeton, NJ, and accepted by the Editorial Board November 20, 2007 (received for review August 2, 2007)

Gene duplication is an important mechanism in the evolution of protein interaction networks. Duplications are followed by the gain and loss of interactions, rewiring the network at some unknown rate. Because rewiring is likely to change the distribution of network motifs within the duplicated interaction set, it should be possible to study network rewiring by tracking the evolution of these motifs. We have developed a mathematical framework that, together with duplication data from comparative genomic and proteomic studies, allows us to infer the connectivity of the preduplication network and the changes in connectivity over time. We focused on the whole-genome duplication (WGD) event in *Saccharomyces cerevisiae*. The model allowed us to predict the frequency of intergene interaction before WGD and the postduplication probabilities of interaction gain and loss. We find that the predicted frequency of self-interactions in the preduplication network is significantly higher than that observed in today's network. This could suggest a structural difference between the modern and ancestral networks, preferential addition or retention of interactions between ohnologs, or selective pressure to preserve duplicates of self-interacting proteins.

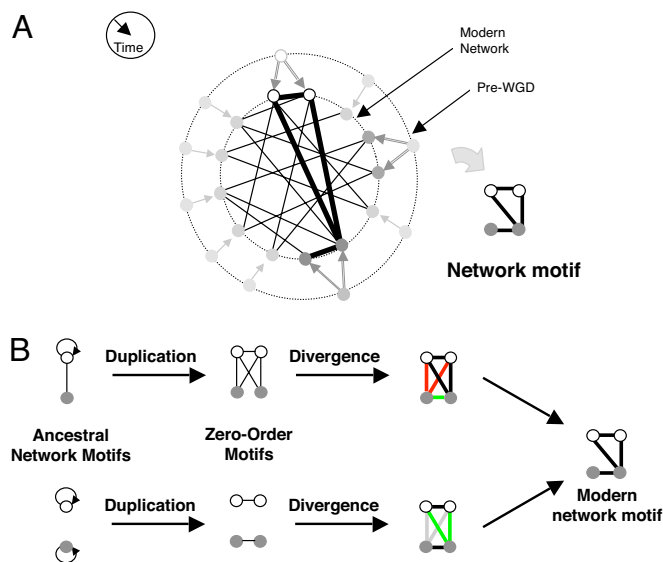
gene duplication | network motifs | self-interacting proteins | whole-genome duplication

Complex biological networks result from the evolutionary growth of simpler networks with fewer components. Gene duplication is thought to be a key mechanism by which networks evolve and new components are added (1–6, 43). These duplication events can act on a single gene, a chromosomal segment, or even a whole genome (1, 7–11). After duplication, the duplicate genes may assume one of several fates, including differentiation of sequence and function, or loss of one of the duplicates (12–17, 44). These outcomes are thought to be affected by genetic factors including redundancy, modularization, and expression dosage (9, 12, 15, 18–22, 45).

Little is known about the rules that govern the modification of gene interactions after a duplication event or the effects of gene interaction on the fate of duplicate genes. Here, we report a mathematical framework for inferring the preduplication connectivity properties of a network and for describing its postduplication dynamics. Our method decomposes a protein interaction network into a vector of network motifs and tracks the evolution of this vector over time. We apply our methodology to the protein interaction network of *Saccharomyces cerevisiae* (23–29), which has undergone a whole-genome duplication (WGD) event, resulting in hundreds of coordinately duplicated gene pairs (ohnologs) (8, 9, 11).

## Results and Discussion

Network motifs are small subgraphs, or interaction patterns, that occur in networks more frequently than would be expected by chance (30). Motifs have been a valuable tool in identifying functional structure in many biological networks including in



**Fig. 1.** Whole-genome duplication (WGD) produces network motifs between ohnolog pairs. (A) The paths genes take through time after a WGD. In most cases only one of the duplicated genes is retained (light gray). Surviving gene duplicate pairs are present as ohnologs in the modern network (white, dark gray). Interactions between any two pairs of ohnologs form a four-node subgraph (network motif) in the proteome. (B) Modern ohnolog motifs are formed through a process of duplication and divergence. Preduplication self-interacting proteins lead to a postduplication interaction between ohnologs. If two ancestral genes interacted, 4 interactions are formed between their pairs of descendants. The duplication step thus yields an initial ohnolog motif (zero-order motifs), which is subsequently modified over time. During the divergence step, interactions might be gained (green) and others are lost (red). Not everything changes: some interactions are retained (black) and other interactions remain absent (gray).

transcriptional, neural, and developmental networks (30, 31). We applied the concept of network motifs to WGD genes in *S. cerevisiae* and analyzed network motifs composed of pairs of ohnologs (namely, motifs of interactions within four proteins, Fig. 1A). There are six possible interactions between any four proteins, hence 64 possible motifs ( $2^6$ ). This number is reduced

Author contributions: A.P., M.B.E., and R.K. designed research; A.P. performed research; A.P., M.K., and R.K. analyzed data; and A.P., M.B.E., M.K., and R.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. L.K. is a guest editor invited by the Editorial Board.

||To whom correspondence should be addressed. E-mail: roy.kishony@hms.harvard.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0707293105/DC1](http://www.pnas.org/cgi/content/full/0707293105/DC1).

© 2008 by The National Academy of Sciences of the USA

**Table 1. Motif distribution in the modern protein interaction network**

Motif class no.	Motif class	No. of motifs present in today's yeast proteome	Modern motif frequency ( $m_{\text{modern}}$ )
1		81,983	$8.15 \times 10^{-1}$
2		17,748	$1.76 \times 10^{-1}$
3		215	$2.13 \times 10^{-3}$
4		925	$9.16 \times 10^{-2}$
5		14	$1.39 \times 10^{-4}$
6		2	$1.98 \times 10^{-5}$
7		93	$9.21 \times 10^{-4}$
8		15	$1.48 \times 10^{-4}$
9		6	$5.94 \times 10^{-5}$
10		0	0
11		16	$1.58 \times 10^{-4}$
12		0	0
13		1	$9.90 \times 10^{-6}$
14		1	$9.90 \times 10^{-6}$
15		0	0
16		4	$3.96 \times 10^{-5}$
17		0	0
18		1	$9.90 \times 10^{-6}$
19		1	$9.90 \times 10^{-6}$

to 19 different motif classes after accounting for the symmetry between the motif's ohnolog pairs and the symmetry of the genes within each ohnolog pair [supporting information (SI) Table 3].

The proteins we considered for our motif analysis are the 450 WGD ohnolog pairs, as listed in Kellis *et al.* (8). Interactions between these proteins are listed in the Database of Interacting Proteins (DIP) (23–29). From these data we determined the modern distribution ( $m_{\text{modern}}$ ) of our 19 motif classes (Table 1). We observe a rich variability in motif prevalences. Even for motifs with the same number of interactions, we observed that frequencies vary across several orders of magnitude, indicating that motif frequencies reflect evolutionary processes rather than

stochastic effects. We then asked how much of the motif distribution observed today could be explained by a neutral model accounting for the evolutionary dynamics of gene duplication after the WGD event.

We developed a model describing protein connectivity within the subnetwork of surviving ohnologs (Fig. 1A) (5, 36). The model consists of two steps: duplication and divergence (Fig. 1B). The duplication step assumes that each protein is duplicated along with all its interactions. Because the two daughter proteins are initially identical to each other, the resulting interaction sets are identical. Accordingly, if a protein was self-interacting, each of its duplicates will be self-interacting, and an interaction will





between the observed abundance  $m_{\text{modern},i}$  and the expected abundance  $m_{\text{expected},i}$ , scaled by the expected number of motifs:

$$E = \sum_i \frac{(m_{\text{modern},i} - m_{\text{expected},i})}{m_{\text{expected},i}}$$

We then minimize  $E$  using the simplex search method (42) implemented by the *fminsearch* function in Matlab, obtaining best-fit values of  $P_i$ ,  $P_{si}$ ,  $P_+$ , and  $P_-$  (see Table 2). The algorithm to estimate the error in the parameters is

- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthonard V, Aiach N, et al. (2006) *Nature* 444:171–178.
- Barabasi AL, Albert R (1999) *Science* 286:509–512.
- Dehal P, Boore JL (2005) *PLoS Biol* 3:e314.
- Ispolatov I, Krapivsky PL, Mazo I, Yuryev A (2005) *New J Phys* 7:145.
- Pastor-Satorras R, Smith E, Sole RV (2003) *J Theor Biol* 222:199–210.
- Hughes AL (1994) *Proc R Soc London Ser B* 256:119–123.
- Wolfe K (2004) *Curr Biol* 14:R392–R394.
- Kellis M, Birren BW, Lander ES (2004) *Nature* 428:617–624.
- Langkjaer RB, Cliften PF, Johnston M, Piskur J (2003) *Nature* 421:848–852.
- Ohno S (1970) *Evolution by Gene Duplication* (Allen and Unwin, London).
- Wolfe KH, Shields DC (1997) *Nature* 387:708–713.
- Conant GC, Wolfe KH (2006) *PLoS Biol* 4:545–554.
- Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) *Mol Syst Biol* 3.
- Kafri R, Bar-Even A, Pilpel Y (2005) *Nat Genet* 37:295–299.
- Lynch M, Force A (2000) *Genetics* 154.
- Tirosh I, Barkai N (2007) *Genome Biol* 8:R50.
- Wagner A (2002) *Mol Biol Evol* 19:1760–1768.
- Papp B, Pal C, Hurst LD (2003) *Nature* 424:194–197.
- Cliften PF, Fulton RS, Wilson RK, Johnston M (2006) *Genetics* 172:863–872.
- Mintseris J, Weng Z (2005) *Proc Natl Acad Sci USA* 102:10930–10935.
- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe K (2006) *Nature* 440:341–345.
- Wapinski I, Pfeffer A, Friedman N, Regev A (2007) *Nature* 449:54–61.
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutillier K, et al. (2002) *Nature* 415:180–183.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) *Proc Natl Acad Sci USA* 98:4569–4574.
- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) *Nucleic Acids Res Database Issue* 32:D449–D451.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. (2000) *Nature* 403:623–627.
- Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) *Nucleic Acids Res* 28:289–291.
- Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D (2001) *Nucleic Acids Res* 29:239–241.
- Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S, Eisenberg D (2002) *Nucleic Acids Res* 30:303–305.
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chkovskii D, Alon U (2002) *Science* 298:824–827.
- Shen-Orr S, Milo R, Mangan S, Alon U (2003) *Nat Genet* 32:64–68.
- DeLuna A, Avendaño A, Riego L, González A (2001) *J Biol Chem* 276:43775–43783.
- Gibson TJ, Spring J (1999) *TIG* 14:46–49.
- Guldner U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) *Nucleic Acids Res* 34:D436–D441.
- Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, et al. (2006) *J Biol* 5:11.11–11.28.
- Wagner A (2003) *Proc R Soc London Ser B* 270:457–466.
- Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) *Nucleic Acids Res* 33:3629–3635.
- Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) *Genome Biol* 8:R51.51–R51.12.
- Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA (2007) *Genome Biol* 8:213.211–218:213.214.
- Britten RJ (2006) *Proc Natl Acad Sci USA* 103:19027–19032.
- Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. (2004) *Nature* 431:946–957.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C* (Cambridge Univ Press, Cambridge, UK).
- Prince VE, Pickett FB (2002) *Not Rev Genet* 3:827–837.
- Wagner A (2001) *Mol Biol Evol* 18:1283–1292.
- Pereira-Leal JB, Teichmann SA (2005) *Genome Res* 15:552–559.
- Marianayagam NJ, Sunde M, Mathews JM (2004) *Trends Biochem Sci* 29:618–625.
- Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hurst LD, Tyers M (2007) *PLoS Biol* 5:e154.
- Yu H, Paccanaro A, Trifonov V, Gerstein M (2006) *Bioinformatics* 22:823–829.
- Musso G, Zhang Z, Emili A (2007) Retention of protein–protein interactions by ancient duplicated gene products in budding yeast. *Trends Genet* 23:266–269.

described in *SI Text*. We tested the model on simulated networks (*SI Text and SI Table 4*) before running on the actual yeast proteome.

**ACKNOWLEDGMENTS.** We acknowledge N. Barkai, M. Brenner, A. DeLuna, E. Lieberman, I. Nachman, I. Wapinski, and K. Wolfe for their advice and helpful discussions and E. Lieberman and R. Milo for critical readings of the manuscript. This work was supported in part by National Institutes of Health Grants GM068763 (to M.B.E.) and R01GM081617 (to R.K.). A.P. was supported by a National Science Foundation Graduate Fellowship and a National Defense Science and Engineering Graduate Fellowship.