ACSESS proceedings
2008

Fourier Analysis of Time Course Microarray Data and its Relevance to Gene Expression Dynamics

Jerry Chen*, Paul Paolini*

*Computational Science Research Center and Department of Biology
San Diego State University, 5500 Campanile Drive, San Diego, CA 92182

CSRC
an interdisciplinary gateway to the future
SAN DIEGO STATE UNIVERSITY

Fourier Analysis of Time Course Microarray Data and its Relevance to Gene Expression Dynamics

Jerry Chen*, Paul Paolini*

*Computational Science Research Center and Department of Biology
San Diego State University, 5500 Campanile Drive, San Diego, CA 92182

Abstract

One of the exciting, ongoing research areas within the fields of bioinformatics and systems biology is the elucidation of gene and protein networks. While there is a large and important effort towards identifying the specific interactions among genes and proteins, there is also a need to understand the dynamics of gene and protein expression over time.

The goal of this study is to use methods from the field of signal processing to understand the dynamics of gene expression. For example, many biological processes such as the cell cycle, cardiac excitation-contraction and the circadian clock are periodic in nature and have underlying genes that are periodically expressed. In the present study, Fourier analysis was used on time-course microarrays to find periodic genes.

Using the Fast Fourier Transform on previously published yeast time-course microarray data, it was found that 313 genes show periodic expression with a spectral magnitude of at least 0.5. Interestingly, there are four dominant periodicities, one of which matches the known yeast cell cycle periodicity. GO annotation and KEGG pathway analysis verify the significant presence of periodic cell cycle genes within the set of 313 genes found. Thus, Fourier analysis is a valuable tool for understanding gene expression dynamics.

1 Introduction

Fourier analysis is a standard methodology used for analyzing the frequency spectrum of signals. It has been extensively applied to many branches of science and engineering, and is built into analysis tools such as *Matlab* and *Mathematica*. The field of molecular biology, however, which has traditionally relied on more qualitative techniques such as Western blotting and immunofluorescence, has only recently seen the use of more quantitative methods. In particular, the technology of microarrays, which measure gene expression on a high-throughput scale, allow the biologist to answer a whole new set of questions. Instead of asking, 'Does this transcription factor increase the expression of this gene?' biologists are able to probe the question, 'Exactly how much does this transcription factor increase the expression of this gene?' Furthermore, since microarrays are high-throughput, they are also able to answer questions such as, 'How many genes does this transcription factor affect?'

With time course microarrays (TCM), which are a set of microarrays taken from different time points in a single experiment, scientists are able to employ Fourier analysis. If the gene expression pattern of a single gene is viewed over several microarrays as a time-domain signal, then Fourier analysis can be used to extract the various frequencies present in that signal. If this is done on a high-throughput scale, all periodic genes can be identified within a genome – i.e., those genes whose expression oscillates at one (or more) frequencies. From there, further experiments can be done to

investigate the molecular mechanisms of these gene expression oscillations – what are the gene and protein networks underlying the dynamics that we see from the microarray data?

In this short paper, it is hoped that the usefulness of techniques such as Fourier analysis for investigating gene expression dynamics and elucidating gene and protein networks can be demonstrated. A brief summary of current work in this research area is presented (Section 2), with a conclusion of results from the analysis of TCM data published by Spellman et al. [1] (Section 3).

2 Background

Living organisms exhibit numerous periodic processes, including cardiac rhythms, calcium oscillations, smooth muscle contraction, membrane potential oscillations, neuronal signals, glycolytic oscillations, cAMP oscillations, insulin secretion, gonadotropic hormone secretion, and the ovarian cycle [2]. In particular, because of their ubiquity and importance in maintaining biological homeostasis, a large effort has been focused on finding those genes regulated by the cell cycle [1, 3-10] and the circadian clock [11-14].

Typically, some form of the Fourier transform has been used to detect genes specific to a biological process such as the cell cycle and circadian clock, where the frequency of the process is known. Most studies have had considerable success using the Fourier transform in detecting periodic genes. For example, Rustici, et al. [4] were able to identify and characterize 407 genes (approximately 8% of the total genome) in fission yeast whose expression was cell-cycle periodic. Spellman, et al. [1], whose data were used in the present analysis, identified 800 putative cell-cycle genes in synchronized *S. Cerevisiae* cultures – although this large number has been challenged by others [6]. Finally, Whitfield, et al. [3] found more than 850 human cell-cycle periodic genes using synchronized HeLa cells. Other studies have used more sophisticated techniques that account for factors such as unevenly spaced sample data and missing time points [15, 16].

The approach used here differs slightly from previous studies. Instead of specifically looking for genes of a particular periodicity, there was an interest in finding *all* significant frequencies present within a time course microarray experiment. In a study identifying genes controlled by a well-characterized frequency (FRQ)-based oscillator in *Neurospora crassa* cultures, Corea et al. [12] also found three genes oscillating at a different frequency, presumably being controlled by a yet uncharacterized oscillator. So, in addition to known periodic biological processes, perhaps other frequencies underlie the expression of some genes, suggesting the presence of uncharacterized oscillatory gene networks.

Of course, in performing such a global search, it is expected to also find genes periodic with the cell cycle or circadian clock, depending on the time interval and sample spacing of the data. Indeed, the detection of such genes in agreement with previous studies serves as a proof-of-concept of this methodology. In the following section, a summary of results is presented from an analysis on the microarray data published by Spellman et al. [1].

3 Methods

Microarray data

Time course microarray data is taken from published material for the Yeast Cell Cycle Analysis Project at Stanford (http://cellcycle-www.stanford.edu), analyzed in Spellman et al. Briefly, gene expression data were obtained every 7 min for 119 min from yeast cells synchronized by α-factor arrest. Expression fold change was measured by comparing synchronized cells to a non-synchronized control group. Intermediate missing time points were filled in using linear interpolation. Genes with two or more consecutive missing time points were discarded from further analysis. Genes with initial (0 min) or final (119 min) missing time points were filled in with the adjacent values (either 7 min or 112 min data value, respectively).

Fourier analysis

Significant frequency components were found using the fast Fourier transform (FFT). Briefly, FFT is a method used to compute the discrete Fourier transform of an evenly-spaced finite length signal. It converts a signal in the time domain into the frequency domain, showing the magnitude of each frequency component present within the signal. The formula is given by

$$X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n] e^{-jk(2\pi/N)n}$$

where N is the signal length and k is the frequency. The FFT function in the MATLAB Signal Processing toolbox was used to find significant frequency components within the microarray data.

Significant frequency components were chosen to be those frequencies $X[k]$ with spectral amplitude $|X[k]| > 0.5$, corresponding to a $\sqrt{2}$ peak-fold change according to the formula

$$|X[k]| = \log_2(\text{peak fold change})$$

where the fold change is the ratio of experimental and control expression values. Future studies could decrease or increase the stringency required for significance.

Gene annotation and pathway analysis

Gene annotation was done using the *S. Cerevisiae* genome database (http://www.yeastgenome.org). Heat maps and significant GO terms were obtained by using analysis tools provided online at Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov, [17]), the Gene Ontology (http://www.geneontology.org/, [18]) and KEGG pathway analysis (http://www.genome.jp/kegg/, [19]).

4 Results

Global gene expression data from time course microarrays were taken every 7 minutes for 119 minutes from yeast synchronized using α-mating factor (for details, see [1]). Fourier analysis using MATLAB extracted genes whose expression signals had significant frequency components. The genes were then identified, organized and annotated.

Identification of significant periodic genes

Genes with frequency components of magnitude greater than 0.5 were found, identified and clustered. Table 1 shows the number of genes found with each periodicity. Note that many genes have more than one frequency component present within their gene expression signal, indicating either multiple underlying cyclic processes or sub-populations of cells oscillating at different frequencies.

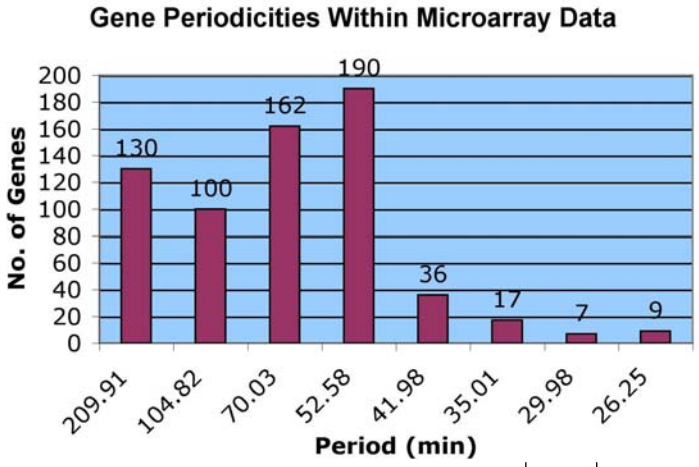

Table 1. Periodic genes present with $|F(\omega)| > 0.5$

The length of the cell cycle depends on many factors, such as cell lineage, availability of nutrients, and the presence or absence of signaling molecules that promote or inhibit growth. Data on *cdc-15*-based synchronized yeast cells from Spellman, et al. [1] suggest that the *S cerevisiae* cell cycle ranges from 50 – 90 minutes. Cho et al. [10] measured a yeast cell cycle of approximately 80 minutes. Based on the results of Table 1, it was hypothesized that the yeast synchronized with α-mating factor had a cell cycle of either 52.58 min or 70.03 min. Although the eventual goal is to use Fourier analysis to uncover novel gene oscillatory networks, the first step is to verify the correctness of the methodology by identifying known oscillatory networks such as the cell cycle. Thus, functional annotation clustering using the DAVID bioinformatics database, Gene Ontology (GO) annotation and KEGG pathway analysis were used to annotate the genes identified and their respective networks.

Annotation of genes using DAVID, GO and KEGG

Genes of the two most abundant periodicities, 52.28 and 70.03 min, were annotated, with the hypothesis of finding an over-representation of processes involved with the cell cycle. Tables 2 and 3 show the 40 most represented biological processes, where the genes are classified according to their GO annotation.

Category	Term	Gene Count	% found	PValue
GOTERM_CC_ALL	site of polarized growth	28	20.29%	1.24E-18
GOTERM_BP_ALL	DNA replication	24	17.39%	2.83E-17
GOTERM_CC_ALL	cell wall (sensu Fungi)	22	15.94%	4.49E-16
GOTERM_CC_ALL	external encapsulating structure	22	15.94%	5.52E-16
GOTERM_CC_ALL	cell wall	22	15.94%	5.52E-16
GOTERM_CC_ALL	bud	25	18.12%	7.95E-16
GOTERM_CC_ALL	chromosome	29	21.01%	1.79E-15
GOTERM_CC_ALL	nuclear chromosome	25	18.12%	6.02E-14
SP_PIR_KEYWORDS	cell cycle	28	20.29%	9.12E-14
GOTERM_CC_ALL	replication fork	14	10.14%	1.47E-13
GOTERM_BP_ALL	DNA-dependent DNA replication	18	13.04%	4.35E-13
SP_PIR_KEYWORDS	dna replication	15	10.87%	7.50E-13
SP_PIR_KEYWORDS	cell cycle control	13	9.42%	2.30E-12
GOTERM_BP_ALL	DNA metabolism	39	28.26%	5.40E-12
GOTERM_BP_ALL	cellular physiological process	127	92.03%	6.95E-12
GOTERM_CC_ALL	bud neck	19	13.77%	8.00E-12
GOTERM_CC_ALL	obsolete cellular component	9	6.52%	9.06E-12
GOTERM_BP_ALL	physiological process	128	92.75%	1.26E-11
GOTERM_BP_ALL	cellular process	127	92.03%	1.92E-11
SP_PIR_KEYWORDS	nucleosome core	7	5.07%	2.38E-10
GOTERM_BP_ALL	reproduction	25	18.12%	5.08E-10
GOTERM_BP_ALL	cell cycle	32	23.19%	6.45E-10
SP_PIR_KEYWORDS	glycoprotein	26	18.84%	1.31E-09
GOTERM_CC_ALL	replication fork (sensu Eukaryota)	9	6.52%	2.02E-09
SP_PIR_KEYWORDS	signal	21	15.22%	2.29E-09
KEGG_PATHWAY	SCE04110:CELL CYCLE	20	14.49%	2.84E-09
GOTERM_BP_ALL	regulation of cyclin dependent protein kinase activity	8	5.80%	2.93E-09
GOTERM_CC_ALL	nucleosome	8	5.80%	3.17E-09
GOTERM_BP_ALL	response to stimulus	36	26.09%	3.18E-09
GOTERM_CC_ALL	cell	133	96.38%	3.49E-09
GOTERM_BP_ALL	regulation of progression through cell cycle	19	13.77%	7.29E-09
GOTERM_BP_ALL	regulation of cell cycle	19	13.77%	7.29E-09
SP_PIR_KEYWORDS	cell division	18	13.04%	8.70E-09
GOTERM_CC_ALL	nuclear nucleosome	7	5.07%	1.08E-08
INTERPRO_NAME	IPR007125:Histone core	7	5.07%	1.38E-08
GOTERM_BP_ALL	cytokinesis	15	10.87%	1.85E-08
GOTERM_BP_ALL	conjugation with cellular fusion	15	10.87%	3.35E-08
GOTERM_BP_ALL	conjugation	15	10.87%	3.35E-08
GOTERM_BP_ALL	sexual reproduction	15	10.87%	3.35E-08

Table 2. GO annotation of genes with periodicity 52.28 min

The descriptions of each column are as follows: *Category*: GO type; *Term*: GO biological process or classification; *Gene Count*: the number of genes of a biological process found; *% found*: the percentage of genes of a biological process found; *PValue*: Fischer Exact p-value.

Category	Term	Gene Count	% found	PValue
GOTERM_CC_ALL	cell wall (sensu Fungi)	26	19.70%	1.75E-21
GOTERM_CC_ALL	cell wall	26	19.70%	2.26E-21
GOTERM_CC_ALL	external encapsulating structure	26	19.70%	2.26E-21
SP_PIR_KEYWORDS	signal	28	21.21%	9.08E-16
GOTERM_CC_ALL	site of polarized growth	23	17.42%	8.90E-14
UP_SEQ_FEATURE	signal peptide	28	21.21%	9.37E-13
GOTERM_CC_ALL	bud	21	15.91%	3.43E-12
GOTERM_CC_ALL	obsolete cellular component	9	6.82%	6.29E-12
SP_PIR_KEYWORDS	glycoprotein	28	21.21%	2.95E-11
GOTERM_BP_ALL	physiological process	122	92.42%	7.62E-11
GOTERM_BP_ALL	cellular process	121	91.67%	1.20E-10
SP_PIR_KEYWORDS	nucleosome core	7	5.30%	2.16E-10
GOTERM_BP_ALL	cellular physiological process	120	90.91%	2.30E-10
GOTERM_BP_ALL	cell cycle	31	23.48%	9.03E-10
GOTERM_BP_ALL	cytokinesis	16	12.12%	1.11E-09
GOTERM_BP_ALL	reproduction	24	18.18%	1.13E-09
GOTERM_CC_ALL	cell	128	96.97%	1.67E-09
GOTERM_CC_ALL	nucleosome	8	6.06%	2.32E-09
SP_PIR_KEYWORDS	gpi-anchor	10	7.58%	5.64E-09
SP_PIR_KEYWORDS	cell cycle	22	16.67%	5.66E-09
GOTERM_BP_ALL	mitotic cell cycle	22	16.67%	7.58E-09
GOTERM_CC_ALL	nuclear nucleosome	7	5.30%	8.21E-09
INTERPRO_NAME	IPR007125:Histone core	7	5.30%	1.08E-08
GOTERM_BP_ALL	cell division	17	12.88%	1.14E-08
GOTERM_BP_ALL	conjugation with cellular fusion	15	11.36%	1.88E-08
GOTERM_BP_ALL	conjugation	15	11.36%	1.88E-08
GOTERM_BP_ALL	sexual reproduction	15	11.36%	1.88E-08
SP_PIR_KEYWORDS	cell cycle control	10	7.58%	2.11E-08
GOTERM_BP_ALL	regulation of cell cycle	18	13.64%	2.36E-08
GOTERM_BP_ALL	regulation of progression through cell cycle	18	13.64%	2.36E-08
GOTERM_CC_ALL	bud neck	15	11.36%	3.05E-08
SP_PIR_KEYWORDS	cell division	17	12.88%	4.69E-08
SP_PIR_KEYWORDS	cell wall	12	9.09%	5.27E-08
SP_PIR_KEYWORDS	lipoprotein	13	9.85%	1.00E-07
GOTERM_BP_ALL	interaction between organisms	15	11.36%	1.03E-07
GOTERM_CC_ALL	extracellular region	8	6.06%	1.99E-07
UP_SEQ_FEATURE	propeptide:Removed in mature form	10	7.58%	4.82E-07
KEGG_PATHWAY	SCE04110:CELL CYCLE	17	12.88%	4.90E-07
INTERPRO_NAME	IPR007124:Histone-fold/TFIID-TAF/NF-Y	7	5.30%	5.70E-07

Table 3. GO annotation of genes with periodicity 70.03 min

Note that, as hypothesized, the cell cycle is significantly represented, as well as many GO processes associated with various stages of the cell cycle (e.g., DNA replication, cytokinesis, cell wall, cell division). Table 4 shows the genes that are represented in the KEGG pathway entry for the cell cycle (entry 'SCE04110: CELL CYCLE' in Table 3).

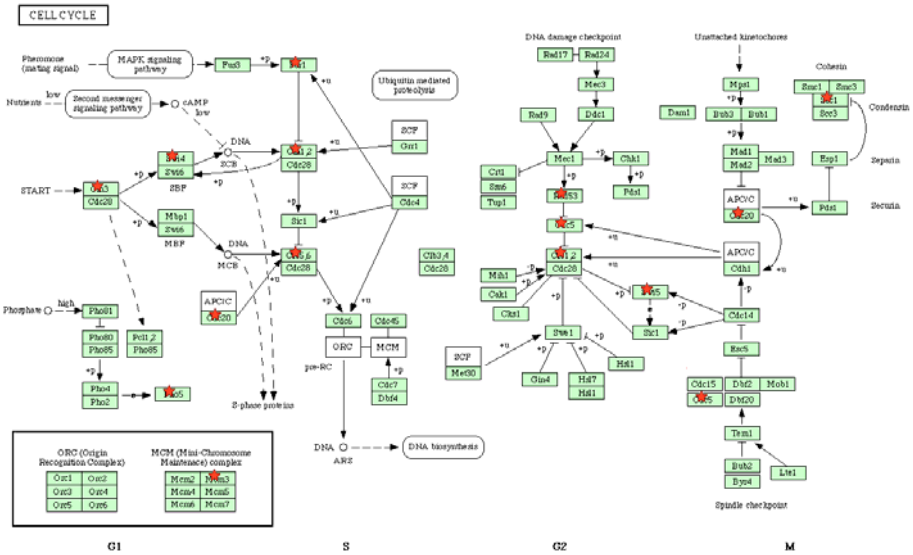

Table 4. KEGG pathway analysis of the yeast cell cycle.

Red stars indicate the positive detection of a significant periodic gene within the data set. Twenty of the genes represented in the KEGG pathway (14.49%, p-value 2.84×10^{-9}) were found. Green boxes are genes that have a detailed gene entry within the DAVID database.

The stringency of fold change required for significance was lowered from $\sqrt{2}$ to 1.25, and subsequent Fourier analysis revealed the presence of 456 and 471 genes with periodicities of 52.58 and 70.03 min, respectively (data not shown). It is likely that even more significant representation of cell cycle processes will be found among these genes.

5 Conclusion

This brief analysis demonstrates the promise of using signal processing techniques such as Fourier analysis as a tool for studying gene expression data from time-course microarrays and discovering genes involved in periodic biological processes. Future directions include (1) annotating and analyzing the genes within this data set underlying other periodicities; (2) analyzing other, more recent time course microarray datasets; (3) developing algorithms that can accurately handle missing, unevenly spaced and short time course data; and (4) using more sophisticated signal processing and statistical methods to better identify periodic genes.

6 Acknowledgments

The author (J.C.) would like to thank Drs. Jose Castillo and Paul Paolini for their support of the project, and Dr. Faramarz Valafar for stimulating discussions that initially motivated the research. This work was supported by NIH NIDDK grant 5K90DK071512-03 "Training for a New Interdisciplinary Workforce" (J.C. is a predoctoral fellow), and by a grant from the California Metabolic Research Foundation (to P.P.).

7 References

1. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. 9(12): p. 3273-97.
2. Ahdesmaki, M., et al., *Robust detection of periodic time series measured from biological systems.* BMC Bioinformatics, 2005. 6: p. 117.
3. Whitfield, M.L., et al., *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Mol Biol Cell, 2002. 13(6): p. 1977-2000.
4. Rustici, G., et al., *Periodic gene expression program of the fission yeast cell cycle.* Nat Genet, 2004. 36(8): p. 809-17.
5. Breeden, L.L., *Periodic transcription: a cycle within a cycle.* Curr Biol, 2003. 13(1): p. R31-8.
6. Shedden, K. and S. Cooper, *Analysis of cell-cycle gene expression in Saccharomyces cerevisiae using microarrays and multiple synchronization methods.* Nucleic Acids Res, 2002. 30(13): p. 2920-9.
7. Marguerat, S., et al., *The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast.* Yeast, 2006. 23(4): p. 261-77.
8. Wichert, S., K. Fokianos, and K. Strimmer, *Identifying periodically expressed transcripts in microarray time series data.* Bioinformatics, 2004. 20(1): p. 5-20.
9. Laub, M.T., et al., *Global analysis of the genetic network controlling a bacterial cell cycle.* Science, 2000. 290(5499): p. 2144-8.
10. Cho, R.J., et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. 2(1): p. 65-73.
11. Panda, S., et al., *Coordinated transcription of key pathways in the mouse by the circadian clock.* Cell, 2002. 109(3): p. 307-20.
12. Correa, A., et al., *Multiple oscillators regulate circadian gene expression in Neurospora.* Proc Natl Acad Sci U S A, 2003. 100(23): p. 13597-602.
13. Bohn, A., et al., *Identification of rhythmic subsystems in the circadian cycle of crassulacean acid metabolism under thermoperiodic perturbations.* Biol Chem, 2003. 384(5): p. 721-8.
14. Stanewsky, R., *Genetic analysis of the circadian system in Drosophila melanogaster and mammals.* J Neurobiol, 2003. 54(1): p. 111-47.
15. Glynn, E.F., J. Chen, and A.R. Mushegian, *Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms.* Bioinformatics, 2006. 22(3): p. 310-6.
16. Liew, A.W., et al., *Spectral estimation in unevenly sampled space of periodically expressed microarray time series data.* BMC Bioinformatics, 2007. 8: p. 137.
17. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. 4(5): p. P3.
18. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. 25(1): p. 25-9.
19. Kanehisa, M. and S. Goto, *KEGG: Kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. 28(1): p. 27-30.