A Maximum Entropy Algorithm for Rhythmic Analysis of Genome-Wide Expression Patterns

Christopher James Langmead[*]

C. Robertson McClung[†]

Bruce Randall Donald [*,‡,†,§,¶]

Abstract

We introduce a maximum entropy-based analysis technique for extracting and characterizing rhythmic expression profiles from DNA microarray hybridization data. These patterns are clues to discovering genes implicated in cell-cycle, circadian, and other periodic biological processes. The algorithm, implemented in a program called ENRAGE *(Entropy-based Rhythmic Analysis of Gene Expression), treats the task of estimating an expression profile's periodicity and phase as a simultaneous bicriterion optimization problem. Specifically, a frequency domain spectrum is reconstructed from a time-series of gene expression data, subject to two constraints: (a) the likelihood of the spectrum and (b) the Shannon entropy of the reconstructed spectrum. Unlike Fourier-based spectral analysis, maximum entropy spectral reconstruction is well suited to signals of the type generated in DNA microarray experiments. Our algorithm is optimal, running in linear time in the number of expression profiles. Moreover, an implementation of our algorithm runs an order of magnitude faster than previous methods. Finally, we demonstrate that* ENRAGE *is superior to other methods at identifying and characterizing periodic expression profiles on both synthetic and actual DNA microarray hybridization data.*

1 Introduction

Certain biological processes are periodic. The cell-cycle and circadian clock, for example, repeat at well defined and reliable intervals. Biologists have shown that the expression patterns of many genes associated these periodic biological processes are themselves rhythmic. Conversely, the expression profiles of genes associated with aperiodic biological processes (e.g., tissue repair) are not rhythmic. The functional significance of previously uncharacterized genes, therefore, may be inferred if they exhibit rhythmic patterns of expression synchronized to some ongoing biological process.

DNA microarray experiments are an effective tool for identifying rhythmic genes when a time-series of expression levels are collected. Unlike Northern blots and PCR, which study one gene at a time, time-series experiments using DNA microarrays reveal the expression patterns of entire genomes. This allows chronobiologists to assign putative functional properties to large numbers of genes based on the results of a single experiment. However, the large volume of data generated by hybridization experiments makes manual inspection of individual expression profiles impractical. Separating the subset of genes whose expression profiles are rhythmic from the thousands, or tens of thousands that are not, requires computer assistance. Ideally, the algorithms for analyzing microarray data should be efficient and have well-understood performance guarantees.

We have designed and implemented an algorithm to identify and characterize the spectral properties of gene expression profiles. Our approach specifically addresses the limitations of Fourier analysis on short time series data — the kind typically generated in microarray experiments. Furthermore, our algorithm is optimal, scaling linearly with the number of gene expression profiles.

The identification of rhythmic genes from microarray data may be achieved by computing the frequency-domain

[*] Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[†] Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.

[‡] Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[§] Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.

[¶] Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: `brd@cs.dartmouth.edu`

Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'02)
0-7695-1653-X/02 $17.00 © 2002 IEEE

spectrum of each expression profile. Rhythmic genes will yield spectra that have a single, well-defined peak in frequency space. In contrast, the spectra of non-periodic profiles will be (relatively) flat. One might think that traditional spectral techniques, like the Fourier transform, are appropriate for obtaining such spectra. However, due to the cost of generating each time point, microarray time-series typically have very few data points. A typical experiment might have only twelve data points and some have as few as four. Traditional spectral techniques, including the Fourier transform, do not perform well on short time series. The resulting spectra do not have adequate resolution for identifying and characterizing oscillating genes [22]. The maximum entropy method (MEM) [29, 26, 5, 24, 32, 3] of spectrum reconstruction is *not* a traditional spectral technique. The MEM *is* well-suited for short time series. We show it is capable of generating smooth, high-resolution spectra from gene expression profiles.

The use of the MEM represents a significant departure from previous techniques for identifying periodic gene expression profiles from microarray data. The dominant methods for identifying rhythmic genes (e.g., [13, 17, 14]) treat the task as either a clustering or pattern recognition problem. Previous algorithms that use hierarchical clustering (e.g., [13]) run in time $O(n^2 l)$, where n is the number of genes represented in the microarray data and l is the number of time-series points. Other algorithms (e.g., [17]) that estimate both the frequency and phase of gene expression profiles using pattern recognition run in time $O(nmp\, l \log l)$, where m is the frequency resolution, and p is the phase resolution. These methods can take up to a week of wall-clock CPU time to analyze data from a single gene chip experiment.

Recently, a method [22] has been described that is both efficient and uses a true mathematical metric to measure morphological similarity. That algorithm runs in time $O(nml^2 + nl^3 \alpha(l) \log l)$, where α is the extremely slow-growing inverse of Ackerman's function. [22] has been shown to be more accurate at estimating frequencies and phases than non-metric based techniques.

In contrast, the algorithm presented below runs in time $O(nl \log l)$. In all cases, l may be treated as a small constant, since in today's technology, l is never more than a small constant $l_{max} \ll n$ (for example, typically, $l \leq 24$, and $n \approx 15,000$ — See Table 1). This simplification obtains complexity bounds of $O(n^2)$ [13], $O(nmp)$ [17] and $O(nm)$ [22] for previous algorithms vs. $O(n)$ for ours. Our algorithm is also, in general, more accurate when determining the frequency and phase of rhythmic profiles than any other previous method.

From a complexity-theoretic point of view, [17] provides a brute-force $O(nmp)$ time algorithm that takes a week to run. [22] uses a phase-independent method to eliminate all complexity dependence on the phase resolution p, yielding an $O(nm)$ algorithm that runs in 2 hours on a set of $10,000$ expression profiles. Our current paper employs the maximum entropy method to eliminate all complexity dependance on frequency resolution m, resulting in an $O(n)$ algorithm that is theoretically optimal, and runs in 22 seconds on the same data set on a Pentium-class workstation.

Our chief contributions are as follows:

1. Elucidation of the limitations of the Fourier transform for the analysis of periodic gene profiles,
2. The use of the maximum entropy method for spectral reconstruction of DNA microarray time-series data,
3. An optimal algorithm for identifying and characterizing rhythmic expression profiles,
4. Unlike previous algorithms, our method has *no* complexity dependance on frequency resolution,
5. Testing our methods on publicly available gene expression data and a comparison of the results to previous methods, and
6. A controlled study of how a variety of methods, including our own, perform in a controlled experiment where signal-to-noise ratio, sample-rate and signal length are varied in synthetic data sets.

Our paper describes an implementation of Maximum Entropy Spectral Analysis (MESA) for rhythmic analysis of genome-wide expression patterns. To our knowledge, this is the first application of MESA to gene-expression microarray time-series data. However, MESA has been previously applied to literally hundreds of problems in the physical and biological sciences (e.g., cf. Dowse and coworkers, who apply MESA to behavioral time-series [11, 12, 23]). The focus of this paper is the improvement in running time and accuracy conferred by MESA, over previous algorithms, for time-series microarray analysis.

1.1 Organization of paper

We begin, in Section 2, with a review of the relevant biology and a summary of three publicly available DNA microarray hybridization time-series data sets. Section 3 categorizes existing techniques for extracting rhythmic profiles from microarray data, including a discussion of their limitations and computational complexity. In section 4, we detail our method and analyze its computational complexity. Section 5 presents the results of the application of ENRAGE to simulated and real biological data. Finally, section 6 discusses these results.

2 Background

There are many examples of DNA microarray time-series experiments in the literature (e.g., [8, 9, 16, 25, 27, 17, 30, 20, 35, 31]). In many of these experiments, the primarily goal was to identify genes whose expression patterns were periodic over the length of the experiment. For example, cell-cycle regulated (e.g., [8, 30]) and circadian (e.g., [25, 27, 16, 9, 17, 35, 31]) genes have been identified from their expression profiles in hybridization experiments.

Several research labs have made their raw data available to the public via [1] facilitating the development of improved techniques. The Davis lab at Stanford has released the yeast data presented in [8] on the CDC28 mutant of yeast. The Botstein lab has released the data from their yeast experiment on the CDC15 mutant of yeast presented in [30]. The Rosbash lab at Brandeis has recently released the data from their circadian experiment on *Drosophila* presented in [25]. In this section we briefly summarize the biological background relevant to these data sets.

2.1 Yeast and *Drosophila* Data sets

The CDC15, and CDC28 experiments were designed to identify cell-cycle regulated genes in yeast (*Saccharomyces cerevisiae*). The eukaryotic cell-cycle is the 4 stage process by which a single cell replicates into two daughter cells. This process takes about 90 minutes in yeast. The authors of the CDC15 and CDC28 experiments were looking for uncharacterized genes whose expression profiles were periodic with 90 minute wavelengths.

The *Drosophila* data set was generated as part of an experiment to identify circadian and circadian-regulated genes. Circadian rhythms are biological processes that are synchronized to the diurnal cycling of light and dark. The goal of circadian microarray experiments is to find the genes associated with the circadian clock. Table 1 details the content of the CDC15, CDC28 and *Drosophila* data sets.

3 Prior Work

A variety of techniques have been developed to extract the rhythmic genes from microarray data sets. The techniques fall into two categories: *spectral* and *pattern matching*- based analyses. In this section we discuss each type, citing specific examples.

3.1 Traditional Spectral Techniques

The Fourier Transform is a standard tool for detecting periodicities in discretized signals. The limitations of the Fourier Transform are well understood. The range of detectable frequencies within a signal, and the resolution to which they can be resolved are particularly relevant to DNA microarray data. Unfortunately, the frequency resolution obtainable on short time series, such as those generated in typical microarray experiments, is often not adequate for resolving periodicities of interest [22]. The sampling rate and sampling interval determine both the range of detectable frequencies and the resolution to which individual frequencies can be resolved. Shorter time series yield spectra that are difficult to interpret. For example, [22] proves that the frequency resolution of the CDC15 data set is 17.9 minutes. In other words, the wavelengths of two periodic functions must differ by at least 17.9 minutes in order to be well-resolved (distinguishable) by the Fourier Transform. This is a fairly coarse resolution given the goal of finding genes with 90-minute wavelengths. Indeed, [30] reports that the Fourier Transform was unstable on the CDC15 data set, and resorted to a hybrid approach including non-spectral methods to estimate frequencies. Most researchers have subsequently used non-spectral methods, described in the next section, to analyze their data.

The size limitations on the data sets are typically not biological but rather financial. Individual microarray chips can cost hundreds of dollars. A per-chip processing fee, also in the hundreds of dollars, is typically charged. Finally, it is necessary replicate the experiment, ideally three or more times, with new samples and chips in order to get statistically accurate estimates of mRNA expression levels. Hence, a DNA microarray time-series experiment can easily cost tens of thousands of dollars.

3.2 Pattern Matching Techniques

Pattern matching algorithms for gene expression analysis take as input the recorded data, a set of models (patterns) with known properties and a method for computing the similarity between a model and an expression profile. For example, the models are typically (a) known periodic gene expression profiles or (b) ideal, synthetic sinusoids. These algorithms assign each gene the properties of the model to which it is most similar.

An important distinction among pattern matching methods is which similarity measurement is used. The choice of similarity measurement affects both the complexity and accuracy of the resulting algorithm. Most clustering algorithms for gene expression profiles use some variation of the *correlation coefficient*. Unfortunately, the correlation coefficient is not a particularly good estimator of shape similarity. Consider, for example, two sine waves of the same frequency that differ in phase by 90°. The correlation coefficient of those two curves is 0, indicating that they are not similar. Nonetheless, the two shapes have a lot in common. Furthermore, the correlation coefficient violates the *triangle inequality* [22] and is therefore not a mathematical metric. The triangle inequality requires that given a distance metric d for comparing expression profiles, for any three expression profiles X, Y, and Z, $d(X,Y) + d(Y,Z) \geq d(X,Z)$.

Experiment	Organism	Δt (minutes)	# samples	# periods	# genes
CDC15 [30]	*S. cerevisiae*	10/20	24	3.2	6178
CDC28 [8]	*S. cerevisiae*	10	17	1.8	6220
Drosophila [25]	*D. melanogaster*	240	6	0.83	14,010

Table 1. CDC15, CDC28 and *Drosophila* data sets. Δt indicates the time period between successive time points. If there is more than one Δt listed, then the data was non-linearly sampled using a combination of the specified times. # periods indicates the number of cell-cycle periods that fit within the duration of the sample interval.

As argued in [10, 22], the triangle inequality is of particular importance, because it guarantees that if several model expression profiles are similar to a given data expression profile, then these model profiles also must be similar to one another. Under the correlation coefficient, however, it is possible for two highly dissimilar model profiles to be similar to the same data profile for one gene. This is highly counterintuitive. Before [22], few previous methods used a similarity measurement that was a true mathematical metric.

In summary, there are a number of problems with the existing approaches for detecting and characterizing rhythmic genes in microarray time-series data. Traditional spectral methods, like the Fourier transform, are not appropriate because a typical microarray experiment generates relatively short time-series. Model-based techniques have had more success on gene expression data but at the expense of computational complexity.

The algorithm presented in [22] addresses many of the issues associated with model-based analyses. [22] uses a true metric, the Hausdorff distance [18, 10], when comparing expression profiles. A corresponding improvement in accuracy was reported. [22] also gives the best complexity bound to date ($O(nm)$ where n is the number of genes and m is the frequency resolution), improving upon the complexity bound of $O(nmp)$ in [17]. This improvement is obtained through the use of the *autocorrelation* function to enable a phase-independent search of frequency-space.

The method presented below is very different. The maximum entropy method is neither a traditional spectral technique nor a pattern matching-based technique. In the MEM, frequency and phase estimates are obtained from frequency-domain spectra. The crucial difference between the MEM and a traditional spectral technique is *how* those spectra are obtained. Spectra are obtained indirectly from the data through the use of statistics. Our method works on massively parallel data sets ($n \approx 15,000,\ l \leq 24$) with time-series as short as $l = 6$ points. Finally, our MEM algorithm runs in time $O(n)$, which is optimal. We discuss the method in the next section.

4 Maximum Entropy Spectral Reconstruction

The Fourier transform produces a spectrum directly from the data. The MEM produces a spectrum that is *consistent* with the data. The MEM treats spectral analysis as an inverse problem, that of recovering the spectrum F of the 'true' expression profile, H. H cannot be observed directly and must be inferred from the recorded data D. The data D are assumed to be contaminated by noise. Thus, $D = H + \epsilon$ and $F = \mathcal{F}(H)$ where ϵ is noise and $\mathcal{F}$ is the Fourier operator. H, D, and ϵ are all represented as scalar functions of time. In the MEM formalism, H is called the *model* and the MEM's job is to fit the best model to the data D.

The MEM algorithm strikes a balance between a goodness-of-fit criterion (realized via a maximum-likelihood formalism) and the Shannon entropy of H. By maximizing entropy, one minimizes the bias of the model H. That is, one does not impose correlations that are not supported by the recorded data D. This is perhaps the most beneficial property of the maximum entropy method. This same property also guarantees that F, the spectrum of H, will be smooth, which is also beneficial in the context of frequency analysis.

H is a variable in the MEM. We desire an H that maximizes the conditional probability $P(H|D)$. This can be rewritten, according to Bayes' formula, as

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}. \tag{1}$$

For purposes of maximization, P(D) can be ignored. We can therefore use

$$P(H|D) \propto P(H)P(D|H). \tag{2}$$

$P(D|H)$ is a goodness-of-fit criterion and, when maximal, gives the *maximum likelihood* solution for H. When assuming Gaussian noise, it can be shown that the maximum likelihood solution, $\mathcal{L}(H, D)$, for D and a specific choice of H is,

$$\mathcal{L}(H, D) = \frac{-\chi^2(H, D)}{2} \tag{3}$$

where

$$\chi^2(H, D) = \sum_{j=1}^{k} |H_j - D_j|^2. \tag{4}$$

Here k is the length of the expression profile. H_j is the jth time point in the model H and D_j is the jth time point in the data D. The probability $P(D|H)$ is obtained by exponentiating[1] $\mathcal{L}(H, D)$:

$$P(D|H) = \exp\left(\frac{-\chi^2(H, D)}{2}\right). \tag{5}$$

The MEM ties $P(H)$ to the Shannon entropy of the spectrum F of H. F is treated as a probability density function (PDF). The Shannon entropy of a PDF A is

$$S(A) = -\sum_{i \in A} p_i \ln p_i, \tag{6}$$

where p_i is the probability of the ith event in A.

When we treat the spectrum F as a PDF, the amplitude $F(\omega)$ of the spectrum at frequency ω, is interpreted as a probability. Thus,

$$S(F) = -\sum_{\omega} F(\omega) \ln F(\omega). \tag{7}$$

Futhermore, because the Fourier operator is an orthonormal change of basis, $S(H) = S(F)$. $P(H)$ is then obtained by exponentiating $S(F)$:

$$P(H) = \exp(S(F)). \tag{8}$$

To maximize Eq. (2), we substitute Eq. (5) for $P(D|H)$ and Eq. (8) for $P(H)$. The MEM then finds the model H_{opt} that maximizes Eq. (2). H_{opt} is the model with maximum entropy among those that are most likely given D. It can be shown that both $\chi^2(H, D)$ and $S(H)$ are convex as H varies. Therefore, there is an analytical solution for finding the H that maximizes Eq. (2) (see Sec. 4.1). For more information on the maximum entropy method, the reader is directed to [29, 26].

As previously noted, $F = \mathcal{F}(H_{opt})$. Our algorithm examines the MEM solution spectrum F_G for each gene G in the data set and determines whether or not the gene is rhythmic. This is accomplished by finding the largest peak in F_G. If the amplitude of that peak is above a chosen threshold, the gene is considered rhythmic. In our experiments we used a threshold of 0 db. The phases and the frequencies of rhythmic genes are obtained directly from each spectrum F_G. Figure 1 demonstrates the difference in quality of spectra generated using the Fourier Transform vs. the MEM.

4.1 Algorithmic Complexity

A number of different numerical techniques have been developed to compute the MEM reconstruction efficiently (e.g., [5, 24, 32, 3]). These techniques take advantage of a mathematical equivalence between F and the frequency response of an all-pole filter derived from the autocorrelation of D. The coefficients of that filter can be computed by solving a set of linear equations using a symmetric Toeplitz matrix built from the autocorrelation coefficients of D [26]. In our implementation we used the Yule-Walker method [32]. The complexity of that algorithm is $O(l \log l)$.

The algorithmic complexity of our algorithm is as follows. For a set of n gene expression profiles, the time to recover all n maximum entropy spectra is $O(nl \log l)$. This improves upon the previous approaches reviewed in Secs. 1 and 3, such as the $O(n^2 l)$ algorithm of [13], the $O(nmp\, l \log l)$ algorithm of [17], and the $O(nml^2 + nl^3\alpha(l) \log l)$ algorithm of [22], where m is the frequency resolution, p is the phase resolution, and α is the inverse of Ackerman's function. Parameters m and p are eliminated in the MEM approach; they can be large—in [17], $m = 1,000$ and $p = 101$.

We argued in Sec. 1 that because l is always small, it can be treated as a constant. The size of a spectrum is $O(l)$. Therefore, the cost of finding the largest peak within the spectrum is constant. Consequently, estimating the frequency and phase of each gene is completed in constant time. Hence, our algorithm runs in $O(n)$ time, which is optimal. That is, treating l as $O(1)$ obtains complexity bounds of $O(n^2)$ [13], $O(nmp)$ [17] and $O(nm)$ [22] for previous algorithms vs. $O(n)$ for ENRAGE. From a practical point of view, the brute-force $O(nmp)$-time algorithm of [17] takes a week to run. RAGE [22] provides an $O(nm)$ algorithm that runs in 2 hours on a set of $10,000$ expression profiles. Our MEM algorithm (ENRAGE) runs in 22 seconds on the same data set on a Pentium-class workstation.

4.2 Non-linearly Sampled Data

Both the Fourier transform and the Yule-Walker method for computing maximum entropy spectral reconstructions assume that the data have been linearly sampled. However, many microarray time-series data sets (e.g., [30, 20]) are not linearly sampled. We note that there exist alternative methods for computing the maximum entropy spectrum on non-linearly sampled data (e.g., [29]). These methods are iterative in nature and are, therefore, more complex algorithmically. However these methods do have the advantage of being able to take advantage of *all* the data, and not just a linearly spaced subset of the data.

[1] Formally, Eqs. (5) and (8) must be scaled by their respective partition functions to be true probabilities. However, for the purposes of maximizing Eq. (2), scaling isn't necessary.

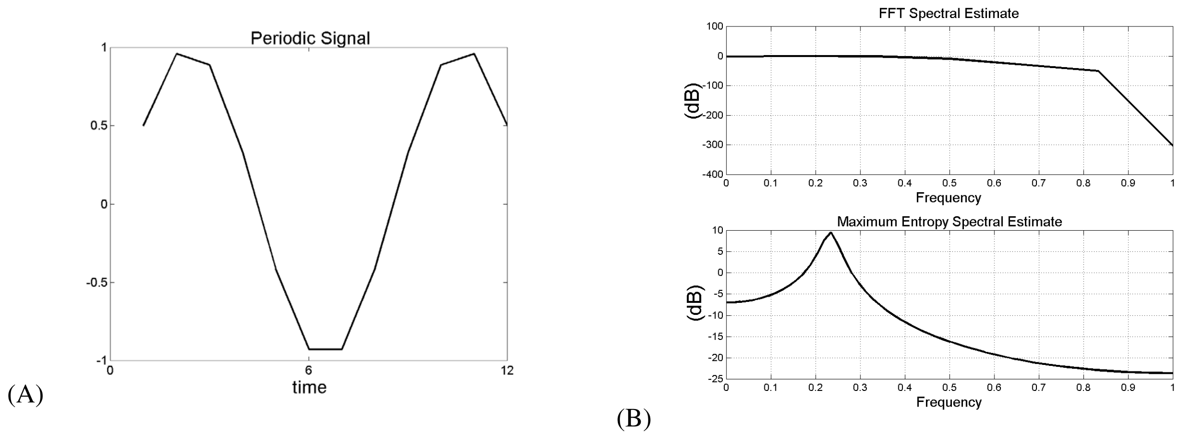

Figure 1. (A) A 12-point periodic signal. (B) Upper panel: The spectral analysis of the signal in panel (A) using the Fourier Transform. Note that the spectrum is essentially flat, making frequency and phase estimation very difficult. Lower panel: The spectral analysis of the signal in panel (A) using the MEM. Here, there is a clearly defined peak in the correct location, making frequency estimation straightforward.

5 Results

ENRAGE has been applied to both synthetic and real microarray data. A series of controlled experiments with synthetic data was designed to compare the performance of ENRAGE, RAGE [22] and a Fourier-based technique. The Fourier-based method is identical to ENRAGE with one exception: it computes the spectrum of the synthetic signals using the Fourier Transform and not using the MEM. Synthetic data sets, each consisting of 10,000 synthetic expression profiles were generated. Each data set contained 5,000 periodic signals (with random phases and frequencies) and 5,000 non-periodic signals. The tasks were to (a) separate the periodic signals from the non-periodic ones and, (b) estimate the frequency and phase of the periodic signal.

In the first experiment, the test variable was the signal to noise ratio (SNR) of the 5,000 periodic signals. 7 data sets were constructed. In the first data set, the periodic signals had no added noise. The SNR of the remaining data sets were 20:1, 10:1, 6.7:1, 3.3:1, 2.2:1, and 1.7:1. ENRAGE had fewer false positives and false negatives than either RAGE or the Fourier method. Table 2 shows the results of the experiment with the data set with a 3.3:1 SNR.

In two additional experiments, ENRAGE's performance under varying sample rates and sampling intervals was studied. In all cases, ENRAGE outperforms the other programs. Furthermore, ENRAGE's accuracy increased with either higher sample rates or increased signal length. Figure 2 summarizes ENRAGE's accuracy at estimating frequency under varying amounts of noise and SNR.

ENRAGE's performance was also examined on the microarray data described in Sec. 2.1. The task of any DNA microarray analysis technique is to find a subset of the genes with a specified property (e.g., cell cycle-regulated, circadian, etc.). Of course, unlike the synthetic data sets described above, the notions of false positives and negatives are not well-defined. In some biological systems, there exist genes whose expression profile properties have been well-characterized in the literature. For example, there are 104 known cell cycle-regulated genes in yeast [34, 7, 28, 21, 2, 15, 33, 19, 6]. Therefore, it is possible to evaluate a given method in terms of false negatives when such information is available. However, due to experimental conditions, it possible for one or more of these known genes to have atypical expression profiles. It is always necessary to go back to the data and examine the profiles of genes that one expected to find.

False positives are much more difficult to evaluate. One can examine the data and look for gross errors. It is customary for biologists to use DNA microarrays as highly parallel screens. Genes with unknown function that exhibit rhythmicity are subsequently studied using slower, more traditional assays.

Direct comparison of the accuracy of multiple techniques on real data is, therefore, best done in terms of false negatives against known genes. Beyond the known genes, it is not possible to do quantitative comparisons other than examining which genes the two methods both identify vs. the genes that were uniquely identified by each method. Qualitative assessments, such the number of genes returned as a fraction of the entire genome can also be helpful.

ENRAGE finds 81, or 78% of the 104 known cell cycle-regulated genes in the CDC15 data set and 67, or 64% in the CDC28 data set. It is not uncommon to consider the results

Method	False Positives	False Negatives
ENRAGE	7%	1%
RAGE [22]	42%	5%
FT	53%	33%

Table 2. Summary of the results of three programs run on a synthetic data set with 5,000 non-periodic genes and 5,000 periodic genes. Each signal was 6 points long. The data in this table were obtained using a data set where the 5,000 periodic genes had a signal to noise ratio of 3.3:1.

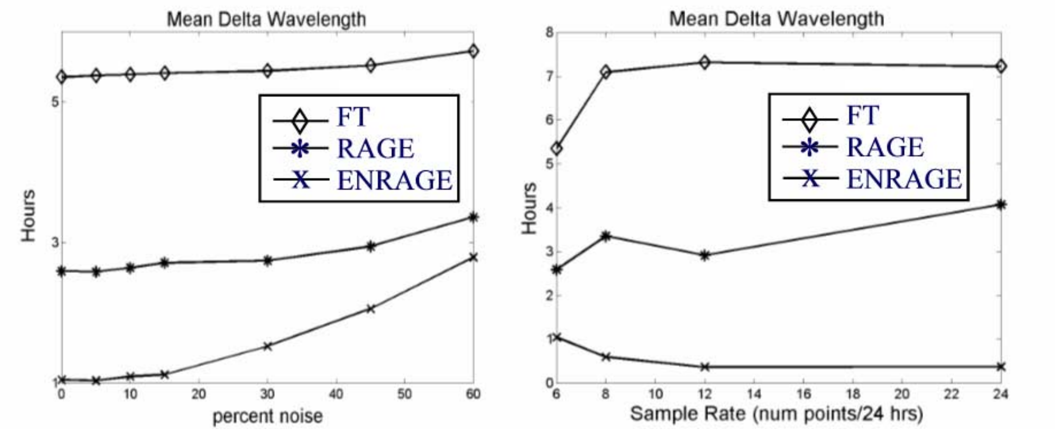

Figure 2. Accuracy of ENRAGE, RAGE [22] and the FT-based estimate under (left panel) decreasing SNR and (right panel) increasing sample rate. Mean Δ Wavelength is the difference between the predicted wavelength and the actual wavelength. ENRAGE has consistently higher accuracy than the other two programs. As the SNR becomes lower, the accuracy of all three programs decreases. The right-hand panel shows that ENRAGE's accuracy increases as the sample rate increases.

from multiple data sets on the same organism when examining genetic data (e.g., [4]). By combining the ENRAGE results on the two data sets, 100, or 96% of the known genes are found. In contrast, [30]'s analysis of the same two data sets yielded only 95, or 91% of the 104 known cell-cycle regulated genes. ENRAGE finds 567 rhythmic genes in both yeast data sets with wavelengths consistent with the length of the cell-cycle in yeast. [30]'s analysis finds 800. It is interesting that ENRAGE's more conservative estimate actually finds more of the 104 known cell cycle-regulated genes.

Finally, ENRAGE was used to analyze the *Drosophila* data set. Of the 14,010 genes in the complete data set, ENRAGE identifies 154 genes as circadian while [25] identifies 134. Both ENRAGE and [25] identify 6 of the known circadian genes (*period, timeless, vrille, clock, cryptochrome, takeout*). In all, 104 genes were identified by both [25] and ENRAGE. Figure 3 shows some of the circadian profiles discovered by ENRAGE.

6 Conclusion

Genome-wide RNA expression time-series experiments are an important source of biological information. The discovery of periodic gene expression profiles is especially useful for the study of rhythmic processes such as the cell cycle and the circadian clock. The sheer volume of data generated by microarray experiments prohibits manual inspection of all the data. Therefore, algorithms for identifying rhythmic genes are needed.

Purely Fourier-based techniques are not yet appropriate for microarray data because the number of time-points in a typical experiment is too small to yield adequate frequency resolution. Model-based techniques are accurate, but are computationally expensive. We have presented a novel technique that is fast, computationally optimal, and substantially more accurate. It gains its efficiency and accuracy through the use of the maximum entropy method of spectrum reconstruction.

Acknowledgments. Thanks to Dr. Hany Farid for his comments and suggestions on this paper. We would like to

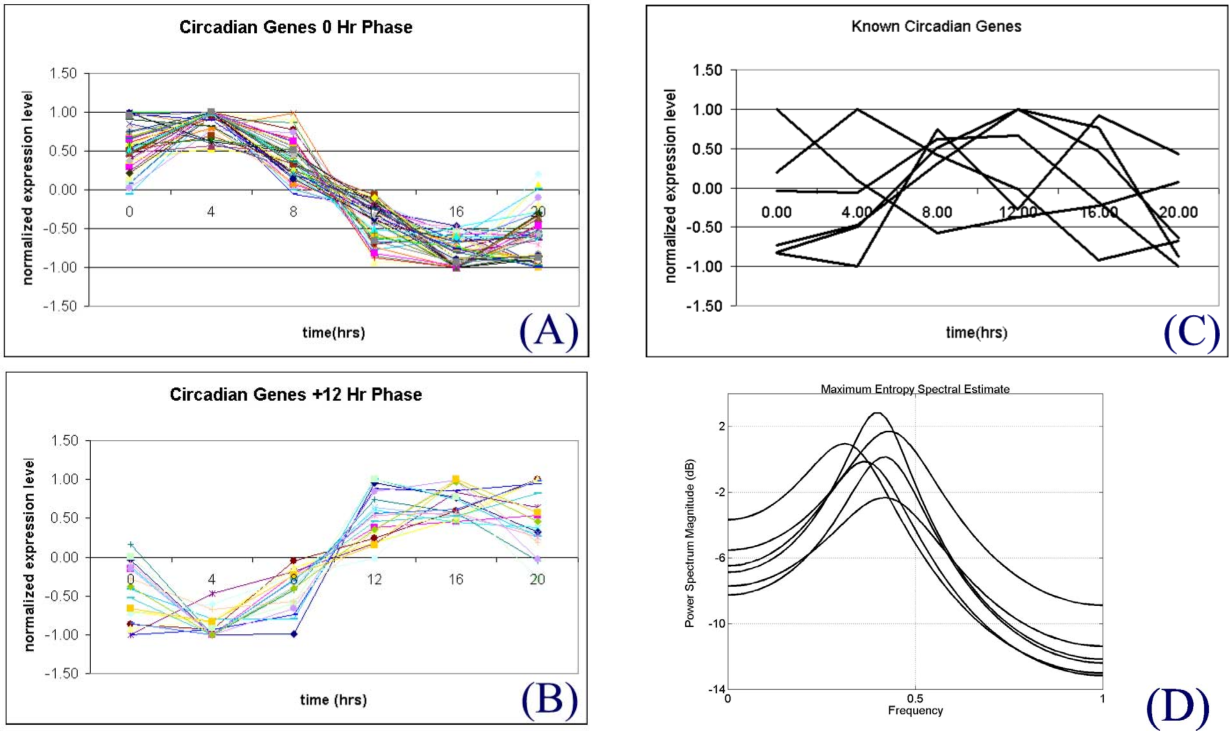

Figure 3. (A,B) Representative clusters from results on *Drosophila* microarray data. Panel A shows the expression patterns of the 47 genes ENRAGE identified as circadian and having a phase offset of 0 hours. Panel B shows the expression patterns of the 20 genes ENRAGE identified as circadian and having a phase offset of 12 hours. (C) Expression profiles of the known circadian genes (*period, timeless, vrille, clock, cryptochrome, takeout*) (D) The 6 ENRAGE spectra for the 6 known circadian genes.

thank Ryan Lilien and all members of Donald Lab for helpful discussions and suggestions. Thanks to Dr. Jeff Hoch for introducing us to the principle of maximum entropy. This work is supported by the following grants to B.R.D.: National Institutes of Health (R01-GM65982), National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068), a Guggenheim Foundation Fellowship, and an equipment grant from Microsoft Research. C.R.M. was supported by National Science Foundation grants MCB-9723482 and MCB-0091008.

References

[1] Stanford Microarray Database. http://genome-www4.stanford.edu/MicroArray/SMD/.

[2] H. Araki, R. K. Hamatake, A. Morrison, A. L. Johnson, L. H. Johnston, and S. A. Cloning DPB3, the gene encoding the third subunit of DNA polymerase II of *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 19:4867–4872, 1991.

[3] I. Barrodale and R. Ericson. An algorithm for least-squares linear prediction and maximum entropy spectral analysis. Part 1 : theory and Part 2 : fortran program. *Geophys*, 45:420–446, 1980.

[4] K. Birnbaum, P. N. Benfey, and D. E. Shasha. cis Element/Transcription Factor Analysis (cis/TF): A Method for Discovering Transcription Factor/cis Element Relationships. *Genome Res*, 11:1567–1573, 2001.

[5] J. Burg. *Maximum entropy spectral analysis*. PhD thesis, Dep. Geophysics Stanford Univ, 1975.

[6] L. H. Caro, G. J. Smits, P. van Egmond, J. W. Chapman, and F. M. Klis. Transcription of multiple cell wall protein-encoding genes in *Saccharomyces cerevisiae* is deferentially regulated during the cell cycle. *FEMS Microbiol. Lett*, 161:345–349, 1998.

[7] J. W. Chapman and J. L. H. The yeast gene, *dbf4*, essential for entry into s phase is cell cycle regulated. *Exp. Cell Res.*, 180:419–428, 1989.

[8] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73, 1998.

[9] A. Claridge-Chang, H. Wijnen, F. Naef, C. Boothroyd, N. Rajewsky, and M. W. Young. Circadian Regulation of Gene Expression Systems in the Drosophila Head. *Neuron*, 32:657–671, 2001.

[10] B. R. Donald, D. Kapur, and J. Mundy. *Symbolic and Numerical Computation for Artificial Intelligence*, chapter 8, "Distance metrics for comparing shapes in the plane," by

D. Huttenlocher and K. Kedem, pages 201–219. Academic Press, Harcourt Jovanovich, London, 1992.

[11] H. B. Dowse, J. C. Hall, and J. M. Ringo. Circadian and ultradian rhythms in period mutants of *Drosophila melanogaster*. *Behavior Genetics*, 17:19–35, 1987.

[12] H. B. Dowse and J. M. Ringo. The Search for Hidden Periodicities in Biological Time Series Revisited. *J. theor. Biol.*, 139:487–515, 1989.

[13] M. Eisen, P. T. Spellman, D. Botstein, and P. O. Brown. Cluster Analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95(25):14863–14868, 1998.

[14] V. Filkov, S. Skiena, and J. Zhi. Analysis Techniques for Microarray Time-Series Data. *Proc. of the 5th Ann.Intl. Conf. on Comput. Biol.*, pages 124–131, 2001.

[15] I. Fitch, C. Dahmann, U. Surana, A. Amon, K. Nasmyth, L. Goetsch, B. Byers, and F. B. Characterization of four B-type cyclin genes of the budding yeast *Saccharomyces cerevisiae*. *Mol. Biol. Cell*, 3:805–818, 1992.

[16] C. Grundschober, F. Delaunay, A. Phlhofer, G. Triqueneaux, V. Laudet, T. Bartfai, and P. Nef. Circadian Regulation of Diverse Gene Products Revealed by mRNA Expression Profiling of Synchronized Fibroblasts. *J. Biol. Chem.*, 276:46751–46758, 2001.

[17] S. Harmer, J. B. Hogenesch, M. Straume, H. S. Chang, B. Han, T. Zhu, X. Wang, J. A. Kreps, and S. A. Kay. Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. *Science*, 290:2110–2113, 2000.

[18] D. P. Huttenlocher and K. Kedem. Computing the minimum Hausdorff distance for point sets under translation. *Proc. 6th ACM Symp. Computational Geom.*, pages 340–349, 1990.

[19] J. C. Igual, A. L. Johnson, and L. H. Johnston. Coordinated regulation of gene expression by the cell cycle transcription factor Swi4 and the protein kinase C MAP kinase pathway for yeast cell integrity. *EMBO J.*, 15:5001–5013, 1996.

[20] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

[21] L. H. Johnston, J. H. White, A. L. Johnson, G. Lucchini, and P. Plevani. Expression of the yeast DNA primase gene, PRI1, is regulated within the mitotic cell cycle and in meiosis. *Mol. Gen. Genet.*, 221:44–48, 1990.

[22] C. J. Langmead, A. K. Yan, C. R. McClung, and B. R. Donald. Phase-independent rhythmic analysis of genome-wide expression patterns. *Proc. of the 6th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB) Washington, D.C., April 18-22)*, pages 205–215, 2002.

[23] J. Levine, P. Funes, H. Dowse, and J. Hall. Signal analysis of behavioral and molecular cycles. *BMC Neurosci*, 3, 2002.

[24] S. Marple. A new autoregressive spectrum analysis algorithm. *IEEE Trans. Acoust. Speech Signal Processing*, ASSP-28:441–454, 1980.

[25] M. J. McDonald and M. Rosbash. Microarray analysis and organization of circadian gene expression in Drosophila. *Cell*, 107:567–578, 2001.

[26] W. Press, B. Flannery, S. Teukolski, and W. Vettering. *Numerical recipes : the art of scientific computing*, chapter 13, Fourier and Spectral Applications, pages 572–575. Cambridge University Press, Cambridge, 1988.

[27] R. Schaffer, J. Landgraf, M. Accerbi, V. Simon, M. Larson, and E. Wisman. Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell*, 13:113–123, 2001.

[28] W. Siede, G. W. Robinson, D. Kalainov, T. Malley, and E. C. Friedberg. Regulation of the RAD2 gene of *Saccharomyces cerevisiae*. *Mol. Microbiol*, 3:1697–1707, 1989.

[29] J. Skilling and S. F. Gull. *Maximum Entropy and Bayesian Methods in Inverse Problems*, chapter "Algorithms and applications" by Skilling, J. and Gull, S. F., pages 83–132. Reidel Publishing Co, Dordrecht, Holland, 1985.

[30] P. Spellman, G. Sherlock, M. Q. Zhang, R. I. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[31] K. Storch, O. Lipan, I. Leykin, N. Viswanathan, F. Davis, W. Wong, and C. Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417:78–82, 2002.

[32] T. Ulrich and T. Bishop. Maximum entropy spectral analysis and autoregressive decomposition. *Rev. Geophys. Space Phys.*, 13:183–200, 1975.

[33] J. Wan, H. Xu, and M. Grunstein. Cdc14 of *Saccharomyces cerevisiae*. Cloning, sequence analysis, and transcription during the cell cycle. *J. Biol. Chem.*, 267:11274–11280, 1992.

[34] J. H. White, S. R. Green, D. G. Barker, L. B. Dumas, and L. H. Johnston. The *cdc8* transcript is cell cycle regulated in yeast and is expressed coordinately with *cdc9* and *cdc21* at a point preceding histone transcription. *Exp. Cell Res.*, 171:223–231, 1987.

[35] K. Yagita, F. Tamanini, G. T. J. van der Horst, and H. Okamura. Molecular Mechanisms of the Biological Clock in Cultured Fibroblasts. *Science*, 292:278–281, 2001.