EPiC Series in Computing
Volume 70, 2020, Pages 139–148
Proceedings of the 12th International Conference on Bioinformatics and Computational Biology

Analysis of Mutation Bias in Shaping Codon Usage Bias and Its Association with Gene Expression Across Specie

Zhixiu Lu[1], Michael A. Gilchrist[2], and Scott Emrich[1]

[1] Department of Electrical Engineering and Computer Science. University of Tennessee, Knoxville,TN 37996, USA
zlu21@vols.utk.edu semrich@utk.edu
[2] Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA
mikeg@utk.edu

Abstract

Codon usage bias has been known to reflect the expression level of a protein-coding gene under the evolutionary theory that selection favors certain synonymous codons. Although measuring the effect of selection in simple organisms such as yeast and *E. coli* has proven to be effective and accurate, codon-based methods perform less well in plants and humans. In this paper, we extend a prior method that incorporates another evolutionary factor, namely mutation bias and its effect on codon usage. Our results indicate that prediction of gene expression is significantly improved under our framework, and suggests that quantification of mutation bias is essential for fully understanding synonymous codon usage. We also propose an improved method, namely MLE-Φ, with much greater computation efficiency and a wider range of applications. An implementation of this method is provided at https://github.com/luzhixiu1996/MLE-Phi.

1 Introduction

Codon usage bias, which refers to using synonymous codons that code for the same amino acid at different rates, has been studied for decades. For example, the Codon Adaptation Index (CAI) relies on relative synonymous codon usage observed in highly expressed genes, and has been effective at predicting gene expression in unicellular microorganisms [11]. Inspired by CAI, tAI goes further and incorporates tRNA gene copy number that exhibits a high and positive correlation with overall rRNA abundance [9]. The underlying assumption behind CAI and tAI is proteins with higher expression contain more optimal codons. Because optimal codons help achieve faster translation with less error, protein-coding genes with a higher ratio of optimal codons likely have experienced more positive selection over time.

Codon usage within multi-cellular organisms with smaller effective population sizes—such as flies, plants and humans—should be less directly affected by selection [9] [14]. To improve prediction performance for all organisms, the Mutation-Selection-Drift balance model was proposed in which selection favors optimal codons and less optimal codons persist due to genetic

Organism	Coding Sequence File Size	Number of Coding Genes	ROC-SEMPPR Run Time	MLE-Φ Run Time
Yeast	5.8 MB	6008	19 Hours	24.32 secs
E. Coli	4.9 MB	4357	17 Hours	23.15 secs
C. Elegans	20 MB	30168	27 Hours	32.22 secs
D. Melanogaster	22 MB	30559	27 Hours	32.94 secs
Arabidopsis	32 MB	48265	29 Hours	34.32 secs

Table 1: **Run time of MLE-Phi vs. ROC-SEMPPR** Run time of ROC-SEMPPR and MLE-Phi for several model organisms. All runs performed on a machine with a i7-6700 CPU and 16 GB of memory.

drift. Codon bias can therefore be thought of a balance between both mutation (e.g., GC content of an organism) and selection (e.g, either high expression or a focus on higher accuracy).

One model that implements this concept is ROC-SEMPPR, which uses a Bayesian Markov chain Monte Carlo (MCMC) to estimate the strength of selection on codon usage [6]. Because this model considers both selective pressure and mutational bias, it can be more comprehensive than models that rely solely on features in highly expressed genes.

Although more inclusive, ROC-SEMPPR's MCMC calculations are also significantly more computationally intensive than most traditional codon usage models. For example, using the current implementation of ROC-SEMPPR requires about 19 hours to process 8.5 Mb of yeast genome data. Codon specific metrics such as CAI and tAI are much faster because they use rely on pre-computed values. For example, the CAI estimate for any given gene sequence is simply the geometric mean of each codon's respective value under the model.

Here, we leverage ideas from ROC-SEMPPR to develop a faster, more flexible codon usage model that also relies on pre-computed values. This new method, which we call MLE-Φ (Maximum Likelihood of Φ), estimates the protein synthesis rate Φ on arbitrary intervals using previously computed ROC-SEMPRR parameters. With this modified Φ estimation framework, we can also predict gene expression at a much finer grain than prior efforts, and we show this using experimental data from several model organisms.

2 Methods

ROC-SEMPPR is capable of calculating codon specific estimates of selection pressure and mutation bias. These estimates have been used to estimate gene expression (Φ) based on this previous equation from [6]:

$$p_i = \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\Phi\right]}{\sum_{j=1}^{n_g} \exp\left[-\Delta M_{j,1} - \Delta\eta_{j,1}\Phi\right]} \tag{1}$$

$\Delta\eta$ is the ROC-SEMPPR measure of relative translation inefficiency for synonymous codons, scaled relative to the preferred codon under selection pressure (Preferred codon has a $\Delta\eta = 0$). In other words, the higher $\Delta\eta$ is, the less efficient the codon is compared to the preferred codon for a specific amino acid. ΔM describes the ratio of the frequencies of one codon relative to the reference under pure mutation; it represents how mutational favored (mutation biased) a codon is relative to the preferred codon. Mutation rates are not always equal, so when there is little selection acting on codon usage (e.g., when gene expression is very low), codon frequencies will be dominated by these more mutation-favored codons as detailed in [6].

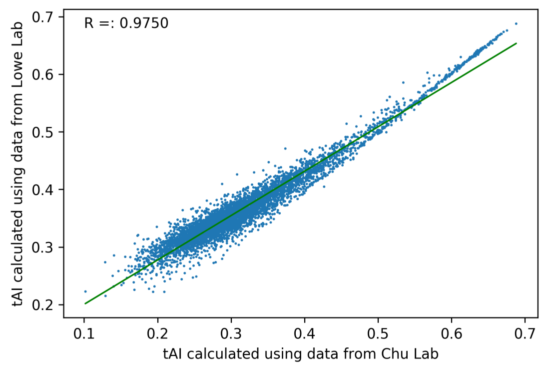

Figure 1: **Correlation between independently derived tAI values: Chu et al. values on the x axis and tRNA counts from Lowe et al. on the y axis. tAI calculations from these two sources yield an almost identical result.**

In the original ROC-SEMPPR model, Φ is calculated by sampling from the posterior distribution from Equation 1 using a Markov chain Monte Carlo (MCMC) approach. Although more explicit, generating a Markov chain for each gene and sampling distributions from it is computationally intensive. Further, the best estimates are often obtained when providing entire genes as input to ensure convergence.

Analysis of codon usage bias in local regions/windows, however, is also important for estimating so-called "translation tempo", i.e., how fast the ribosome translates specific regions of a transcript. To compute Φ for local regions more efficiently, we took an maximum likelihood approach that maximizes the probability in Equation 1 using the revised formula below:

$$\prod_{n}^{n+k} \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\Phi\right]}{\sum_{j=1}^{n_g} \exp\left[-\Delta M_{j,1} - \Delta\eta_{j,1}\Phi\right]} \tag{2}$$

Here n marks the start position of a codon window/interval that spans k codons (when this formula is applied to an entire gene, $n = 0$ and k = gene length / 3). By finding a Φ that maximizes the output probability for this specific window, we can get a effective estimate of Φ much faster, especially since MLE-Φ is optimized by Newton's root approximation method. In experimental studies such as a ribosome footprint count analysis (local translation rates), it has been shown that the ribosome covers about 10 codons in a transcript, suggesting an ideal value for k should be approximately ten for modeling protein translation.

Implementation of MLE-Φ and respective computed values of $\Delta\eta$ and ΔM for several most studied organisms are hosted on **GitHub**, a sample run case is also included.

Experimental analysis

We also consider more complex mechanisms of codon usage bias. In this paper, we first compare our MLE-Φ method with traditional selection based codon usage metrics such as tAI and CAI, using experimental expression measurements from yeast, then shift to more complex organisms to further validate and quantify the effect of mutation bias on codon usage bias.

Specifically, we use the original methods as described in [11] [9] for both tAI and CAI, which estimates tRNA abundance from genomic sequences. To confirm this data source, we

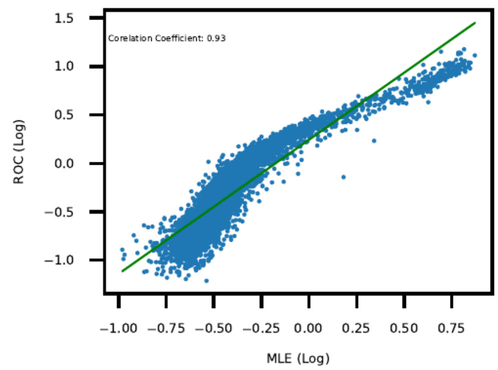

Figure 2: **Correlation between MLE Φ and ROC-SEMPPR Φ** As shown there is a 0.93 Pearson correlation between these two measures, which indicates that our new MLE estimation framework closely corresponds with the calculations from the original ROC-SEMPPR.

found another measurement of tRNA abundances for the yeast s288c strain from the Chu lab [4],

which measures tRNA counts in a cell. We calculated tAI values for all considered genes using both metrics, and values generated using these two sources have a correlation of 0.97 (see Figure 1). For CAI, we chose the top 5% of the highly expressed genes in empirically determined expression data for computing the required model values, and use these with the simple formula first proposed by Sharpe and Li to estimate the expression level of each gene.

3 Results

Run time comparison

Using pre-computed values of $\Delta\eta$ and ΔM to find the maximum likelihood of Φ significantly reduces computation time for estimating Φ. In Table 1 we benchmark both methods using different model organisms.

Note that the original ROC-SEMPPR must be run at least once, so computing the initial values is dominated by the original MCMC calculations. We show, however, that optimized MLE-Φ can produce subsequent estimates for these organisms in only seconds (versus days in some cases). Further, this revised framework can now estimate protein translation rates for more local regions of any given gene (see Methods).

3.1 ROC-SEMPPR vs MLE-Φ estimates

There is little downside in the significant improvement in computation speed as MLE-Φ closely approximates MCMC-based Φ (Figure 2). As expected, the agreement of the two approaches tends to be better for highly expressed genes. This is to be expected as codon usage bias should have stronger detectable effects on genes with higher expression ([13] [7] [8]). Low expression genes correlate less well, in part because they tend to be noisier and harder to measure experimentally ([2] [12]). Even so, there is a clear and strong correlation between the

	MLE-PHI	tAI	CAI
Arava 2003	0.637	0.621	0.643
Sun 2012	0.602	0.600	0.560
Nagalakshimi 2008	0.5214	0.532	0.500
Holstege 1998	0.763	0.710	0.718
Causton 2001	0.688	0.676	0.657

Table 2: **Comparison of three metrics for different yeast data** A comparison between our three considered metrics using previously published yeast mRNA abundances. Based on the Pearson correlation between predictions and empirical gene expression data, all three methods perform similarly in yeast.

Organism	Common Name	Estimated Selection Pressure
Arabidopsis thaliana	Thale Cress	0.24
Caenohardibtis elegans	Roundworm	0.45
Drosophila melanogaster	Fruitfly	0.31
Homo sapiens	Human	0.03
Plasmodium falciparum	Malaria Parasite	0.17
Saccharomyces cerevisiae	Baker's Yeast	0.77
Schizosaccharomyces pombe	Fission Yeast	0.82

Table 3: **Estimation of Selection Pressure in Several Eukaryotes**

original and our new approach with an overall correlation coefficient of 0.93.

3.2 Comparison of tAI, CAI and MLE-Φ Estimates

We tested the performance of Φ estimation using empirical gene expression data relative to both CAI and tAI. This assessment will determine what effects (if any) incorporating mutation bias (ΔM) has on our predictions. We computed these gene expression measurements and then computed their correlation using the same approach as Causton et al. (2001). MLE-Φ's correlation is always higher than CAI for all data, and higher than tAI for 3/5 data sets considered (see Table 2). Combined, this supports using our new Φ estimation framework for predicting gene expression.

3.3 Looking at the effects of other factors affecting expression

Based on the Selection-Mutation-Drift model, more complex organisms with smaller effective populations sizes should be more tolerant of drift and therefore are expected to be less affected by selection pressure. For example, in the original tAI paper the authors estimated the selection pressure on different organisms. Although yeast, considered above, has strong estimated pressure (0.77-0.82), this pressure is only 0.24 in the model organism *Arabidopsis thaliana* and almost non-existent in human (0.03) as shown in Table 3.

Because the more inclusive MLE-Φ model should perform better than CAI and tAI for more complex organisms, we next decided to compare different metrics for organisms under less selection pressure. Although MLE-Φ has the highest overall correlation for all organisms (Table IV), Φ is not always significantly better that tAI and CAI (Z score, $p < 0.01$).

As described previously, measurement (and therefore assessment) is more difficult for genes with lower overall expression ([2] [12]). It is also possible that a given gene may have different expression levels under different conditions/cell types in multicellular organisms. To address this issue, "House keeping" genes have historically been used, which are genes involved in basic

	Sample Size	MLE-Phi	tAI	CAI	Z Score (Phi,tAI)	Z Score (Phi,CAI)
Yeast	518	0.765	0.696	0.726	2.39 (p=0.0168)	1.41 (p=0.1585)
Roundworm	2190	0.606	0.579	0.58	1.38 (p=0.1676)	1.33 (p=0.1835)
Fruitfly	1393	0.546	0.424	0.309	4.22 (p=0)	7.73 (p=0)
Arabidopsis	2756	0.302	0.257	0.106	1.81 (p=0.07)	7.62 (p=0)

Table 4: **Correlation-based comparison of the three considered metrics using the top 5% of highly expressed genes and empirical expression data** Fisher's R-Z transform is used to compute the Z score

cell maintenance that are expected to maintain consistent expression levels irrespective of tissue type, developmental stage, or external signals. Although there are also a few genes such as 16S, tus, rpoD, glyA, dnaB, gyrA, pykA/F, pfkA/B, mdoG and arcA that are widely used, it is difficult to obtain these specific values for the organisms we are studying [5].

To overcome this issue we extend a previously published method from 2007 [5] that used RT-PCR-based abundance estimates to rank genes. By picking genes on the top of the generated rankings, our selections would likely be "housekeeping" gene candidates and, more importantly, for this analysis, have more stable expression levels. Here, rather than using RT-PCR RNA abundance data we create rankings based on RNA-seq expression data from each organism and analyze the top 5% of the highly expressed genes based on these data. As shown in table 4, this approach generates candidates that are less noisy when compared to considering all protein-coding genes. After reconsidering the Pearson correlation coefficients between the considered metrics and prior empirical measurements, our Φ framework still consistently outperforms other methods for all organisms tested.

We also analyzed the difference between correlation coefficients using Fisher's R-Z transform. As shown above, we observe consistently positive Z scores with most comparisons having a corresponding p-value less than 0.05. This further confirms our hypothesis that, by weighting in the effect of mutation bias, Φ-estimation is more comprehensive and therefore a more accurate estimate fo organisms where selection pressure is not the dominant driver of codon usage bias.

3.4 Looking deeper into mutation bias

We have shown above that our MLE estimate of Φ has better accuracy than other traditional codon usage metrics when mutation bias impacts gene expression level estimates.

To confirm that mutation bias is responsible for the observed differences between model predictions, we created rankings for each coding gene in *D. melanogaster* using Φ, tAI, and the prior empirical measurements. We then sorted the genes by the ranking distance differences between tAI and Φ relative to the empirical measurement data. As we move from genes with the highest prediction differences to the lowest, we observe a clear shift in GC content (see Figure 3). This further confirms that mutation bias plays an important role in computational prediction of gene expression, especially for multi-cellular organisms such as *Drosophila*.

3.5 Local vs Global estimates

To ascertain how mutation bias affects so-called "translation tempo", or the rate by which a ribosome transcribes a specific region, we compared local measurements of MLE-Φ and CAI using a window-based analysis (similar to our prior work in [3]), created a ranking of CAI and

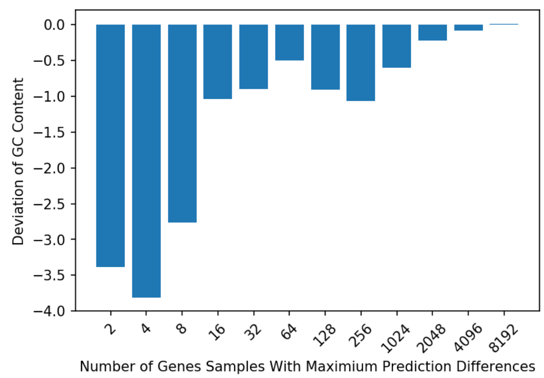

Figure 3: **Shift of GC content across genes with difference levels of prediction differences between using Φ and tAI** The x-axis represents the number of genes with the highest prediction differences between Φ and tAI, samples with a smaller size contain genes with more prediction differences, while the y-axis represents the deviation from sample mean of GC content to population mean calculated from all 11,196 coding genes in *Drosophila*). The observed GC bias decreases as we sample less different predictions between Φ and tAI.

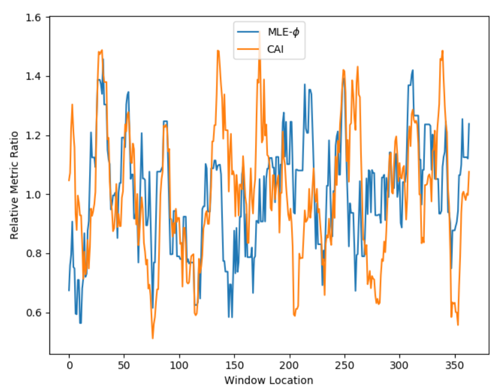

Figure 4: **Relative MLE-Φ and CAI Window Estimation** MLE-Φ and CAI for k=10 codon windows for the ACT1 gene in yeast; values along x axis mark the start codon position of the window, values on the y axis represent the ratio between window metric estimate and whole gene metric estimate. This illustration indicates that although ACT1 is a "housekeeping" gene with consistent global gene expression estimates (difference in ranking $< 1\%$) using different methods, there is visible disagreement in the more local translation rate estimates using these approaches.

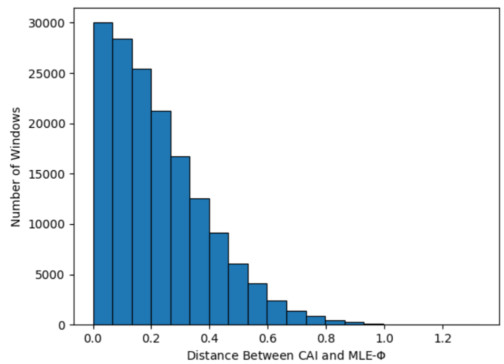

Figure 5: **Distribution of Window Measurements by CAI and MLE-Φ** Figure shows distributions of distance between CAI and MLE-Φ, x label shows the distance if relative metric ratio between MLE-Φ and CAI, values along y axis represent the number of windows (window size of 10 codons) with respective measurement distance.

Φ-based gene expression estimates, and selected a total of 300 genes (5% of roughly 6000 coding genes in yeast) with the least difference in overall expression level predictions.

For an example gene, we present ACT1(YFL039C) that is a house keeping gene ranked in the top 5% of highly expressed by MLE-Φ, CAI and all other methods considered here. As expected, gene-wide predictions of the expression level of ACT1 are very similar; however, we see clear variations in certain local regions (see Figure 4).

This result indicates that local protein translation rate estimates between models can vary, even when global gene expression level predictions are similar. The rationale is CAI and MLE-Φ are global estimates that converge to the same level but do not indicate how fast/slow specific regions are translated. To illustrate this more clearly, we computed the distance between MLE-Φ and CAI for codon windows in a total of 300 (roughly 5%) of the genes described above (Figure 5). Although most genes have similar estimates using CAI and MLE-Phi, there are a number like ACT1 that differ by a substantially. This is a major contribution of this work since there was no local/window-based version of ROC-SEMPPR Φ prior to our developing MLE-Φ as reported here and therefore such differences between global and local estimates has not yet been reported to the best of our knowledge.

4 Discussion

Gene expression is a topic of great interest in biology, and there are a wide range of approaches to model it [1] [15]. For example, prior work has applied probabilistic and machine learning approaches based on microarray data and typically achieve a prediction accuracy between 73% to 79% in yeast [1]. Similar performance is achieved using codon usage bias-based estimates such as CAI and tAI. We extend and improve upon ROC-SEMPPR to develop a new MLE-Φ framework, which allows estimating expression using any arbitrary interval. This allows using codon usage bias to better understand other areas of biological interest such as protein synthesis rates and co-translational protein folding.

Estimation of Φ also provides a more comprehensive interpretation of codon and incorporating mutation bias estimates ΔM from ROC-SEMPPR. We confirm that mutation bias plays an important role in shaping observed codon usage bias. By only selecting top 5% genes that are highly expressed, which is the exact method that underlies CAI and TAI-based estimates, we observe that our new method MLE-Φ is always better. This suggests that incorporating mutation bias into the expression model better predicts the precise expression level of a gene, even in highly expressed genes that are expected to have codon usage dominated by selection. This discovery is most important for more complex organisms like *D. melanogaster*, Arabidopsis and humans.

Significantly, we provide for the first time a framework that can use selection and mutation-based parameters for more localized windows. Prior work, including ours ([3]), have shown that rare codons are evolutionary conserved and some likely help proteins fold by slowing down the ribosome translation complex, a phenomenon called "co-translational folding" in the literature. We show that a number of genes in yeast have the same global estimate but differ greatly in a more local window-based analysis (see Figures 4 and 5). We are currently using our new MLE-Φ framework with ongoing experimental validation of preferred codon usage models for genes known to co-translationally fold (see [10] for details). This will provide biological support that MLE-Φ, which incorporates selection and mutational bias to better predict overall gene expression, also is better able to estimate the "tempo" of the ribosome and aid in downstream protein-focused research.

References

[1] Michael A. Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, 2004.

[2] Joshua S Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10(1):221, 2009.

[3] Julie L. Chaney, Aaron Steele, Rory Carmichael, Anabel Rodriguez, Alicia T. Specht, Kim Ngo, Jun Li, Scott Emrich, and Patricia L. Clark. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Computational Biology*, 13(5), May 2017.

[4] Dominique Chu, David J. Barnes, and Tobias Von Der Haar. The role of trna and ribosome competition in coupling the expression of different mrnas in saccharomyces cerevisiae. *Nucleic Acids Research*, 39(15):6705–6714, Sep 2011.

[5] Eli Eisenberg and Erez Y. Levanon. Corrigendum to: Human housekeeping genes, revisited. *Trends in Genetics*, 30(3):119–120, 2014.

[6] Michael A. Gilchrist, Wei-Chen Chen, Premal Shah, Cedric L. Landerer, and Russell Zaretzki. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone ‡. *Genome Biology and Evolution*, 7(6):1559–1579, 2015.

[7] M. Gouy and C. Gautier. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10(22):7055–7074, 1982.

[8] Ruth Hershberg and Dmitri A. Petrov. Selection on codon bias. *Annual Review of Genetics*, 42(1):287–299, 2008.

[9] M. D. Reis. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32(17):5036—-5044, 2004.

[10] Anabel Rodriguez, Gabriel Wright, Scott Emrich, and Patricia L. Clark. codon usage and its impact on protein folding. *Protein Science*, 27(1):356–362, 2017.

[11] Paul M. Sharp and Wen-Hsiung Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281–1295, 1987.

[12] P. K. Tan. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19):5676–5684, Jan 2003.

[13] Edoardo Trotta. Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Research*, 41(20):9382–9395, 2013.

[14] Edward W.j. Wallace, Edoardo M. Airoldi, and D. Allan Drummond. Estimating selection on synonymous codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30(6):1438–1453, 2013.

[15] Yuan Yuan, Lei Guo, Lei Shen, and Jun S. Liu. Predicting gene expression from sequence: A reexamination. *PLoS Computational Biology*, 3(11), 2007.