Prediction of ribosome footprint profile shapes from transcript sequences

Tzu-Yu Liu\textsuperscript{1,2} and Yun S. Song\textsuperscript{1,2,3,*}

\textsuperscript{1}Department of Mathematics and Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA, \textsuperscript{2}Department of Electrical Engineering and Computer Sciences and \textsuperscript{3}Department of Statistics and Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA

\textsuperscript{*}To whom correspondence should be addressed.

Abstract

Motivation: Ribosome profiling is a useful technique for studying translational dynamics and quantifying protein synthesis. Applications of this technique have shown that ribosomes are not uniformly distributed along mRNA transcripts. Understanding how each transcript-specific distribution arises is important for unraveling the translation mechanism.

Results: Here, we apply kernel smoothing to construct predictive features and build a sparse model to predict the shape of ribosome footprint profiles from transcript sequences alone. Our results on \textit{Saccharomyces cerevisiae} data show that the marginal ribosome densities can be predicted with high accuracy. The proposed novel method has a wide range of applications, including inferring isoform-specific ribosome footprints, designing transcripts with fast translation speeds and discovering unknown modulation during translation.

Availability and implementation: A software package called riboShape is freely available at \url{https://sourceforge.net/projects/riboshape}

Contact: yss@berkeley.edu

1 Introduction

Gene expression is a fundamental biological process consisting of two parts: \textit{transcription} of mRNAs from DNAs and \textit{translation} of proteins from mRNAs. Studying the genome-wide dynamics of both processes is crucial for understanding how cells function and respond to various environmental conditions, thereby giving rise to the impressive complexity of living organisms. Due to technological challenges, translation has not been studied as extensively as transcription, but recent advances in experimental protocol and sequencing technology are providing an unprecedented opportunity to examine the translation process at base-pair resolution.

Ribosome profiling (Ingolia \textit{et al.}, 2009; Ingolia, 2014) applies either translation inhibitors or a flash-freeze protocol to immobilize the ribosomes ‘walking’ along the transcript, and the ribosome-occupied regions (each of length 28–30 nucleotides) can then be extracted and sequenced, providing detailed positional information about ribosomes. Such snapshots of ribosome footprints enable quantitative monitoring and analysis of translational dynamics, making ribosome profiling an important technique for studying protein synthesis. In particular, it has been shown that protein abundance correlates better with ribosome densities than with mRNA abundance (Ingolia \textit{et al.}, 2009).

Various factors may influence the translation mechanism. It has been proposed that the differential usage of synonymous codons, also known as the codon usage bias, is driven by evolutionary forces to facilitate translation speed (Burgess-Brown \textit{et al.}, 2008; Gardin \textit{et al.}, 2010; Shah \textit{et al.}, 2011). This hypothesis is supported by the finding that transcripts with wobble base-pairing encounter slower elongation (Stadler and Fire, 2011), and also by the analysis of tRNA-adaptation index (tAI) (Tuller \textit{et al.}, 2010, 2012; Kudla \textit{et al.}, 2009; Gu \textit{et al.}, 2010; Tuller \textit{et al.}, 2010b, 2011; Zur and Tuller, 2012). Studies have found that initiation is the rate-limiting factor of translation and that ribosome densities tend to be higher near initiation sites in \textit{Saccharomyces cerevisiae} (Ingolia \textit{et al.}, 2009; Shah \textit{et al.}, 2013). Also, positively charged residues on nascent peptides have been found to influence the footprint abundance (Charneski and Hurst, 2013; Lareau \textit{et al.}, 2014; Tuller \textit{et al.}, 2011). We refer the reader to Ingolia (2014) for a review of the major findings from utilizing ribosome profiling.

Although the debate surrounding the main determinant of translational dynamics remains open, the above-mentioned hypotheses have one thing in common; that is, the transcript sequence context plays an important role in governing the efficiency of translation. Hence, investigating how transcript-specific distributions arise and to what extent they depend on the transcript sequence is essential to...
understanding the translation mechanism. Motivated by this observation and the fact that ribosome footprint profiles are generally not uniformly distributed, we consider here the statistical problem of predicting marginal densities of ribosome footprints from the transcript sequence information alone.

We first apply wavelet analysis to decouple global shapes from local patterns of ribosome marginal densities. Then, by building sparse models to predict the marginal densities projected onto the subspace corresponding to each scale and using asymmetric kernel smoothing, we identify the codon features that best associate with a given scale and estimate the extent to which they influence the ribosome densities. Results on *S. cerevisiae* ribosome footprints show that the sequence content is highly predictive of the marginal densities at steady state. The ability to predict the marginal densities based solely on transcript sequences has many potential applications. For example, it can be used to guide the design of transcripts with optimal translational efficiency.

In Section 2, we describe our proposed method, which consists of the ‘A’ site (decoding site where the aminoacyl tRNA arrives) identification, wavelet decomposition, kernel smoothing and sparse regression model. In Section 3, we test the performance of our method on ribosome profiling data of *S. cerevisiae*. We then conclude with a discussion of our findings.

2 Methods

2.1 Ribosome footprint alignment and the ‘A’ site identification

The first step of ribosome profiling is immobilizing the ribosomes on transcripts, which can be accomplished by the application of cycloheximide (CHX), proposed by Ingolia et al. (2009). Then, by nuclease digestion, the fragments of the transcript protected by the ribosomes are retained. After converting these footprints into DNA molecules for deep sequencing, one can measure the abundance of ribosomes in the transcript. This information can serve as a prior in probabilistically mapping reads in the genome.

Wavelet transformation has been widely used for analyzing signals. Its time-frequency localization (Daubechies, 1990) and multiresolution approximation (Donoho, 1995; Donoho and Johnstone, 1997) have been applied to genetic association analyses in high-throughput sequencing (Donoho et al., 1996; Kerkyacharian and Picard, 1992). For wavelet analysis to build a multiresolution approximation of a signal, one can measure the abundance of footprints at each position along a transcript.

We considered four publicly available datasets of ribosome profiling footprints of *S. cerevisiae* treated with CHX, available at the NCBI Gene Expression Omnibus under accession GSE53313, GSE52119, GSE34438 and GSE55400, and published in the recent literature (Albert et al., 2014; Ingolia et al., 2009; McManus et al., 2014). We analyzed the ribosome footprints of wild type *S. cerevisiae* under normal conditions to study the steady-state behavior. The ability to predict the marginal densities based solely on transcript sequences has many potential applications. For example, it can be used to guide the design of transcripts with optimal translational efficiency.

There are three active ribosomal sites: the A (arrival), P (poly peptide) and E (exit) sites, where aminoacyl tRNA arrives for decoding, the poly peptide is created and the uncharged tRNA exits, respectively (Fig. 1). We considered reads of length 28, 29 and 30 nucleotides, and focused on the A sites of these ribosome footprints. For a 28-nucleotide footprint read, the A site typically starts at the 16th nucleotide downstream of the 5' end of the read (Ingolia et al., 2009; Lareau et al., 2014). If that position does not correspond to the first nucleotide of a codon, then, as illustrated in Figure 1, the start position of the A site was adjusted so that it is at either the 15th or the 17th nucleotide, depending on the amount of shift. We applied the same procedure for 29-nucleotide reads. For 30-nucleotide reads, the A site was inferred to start at the 16th, the 17th, or the 18th nucleotide downstream of the 5' end. Marginal probability density functions of the footprints were obtained by normalizing the histograms of the inferred A site footprint counts along each transcript.

We also considered the ribosome footprints obtained using the flash-freeze protocol without CHX pre-treatment (Weinberg et al., 2016). These footprints are publicly available at the NCBI Gene Expression Omnibus under accession GSE53313.

2.2 Wavelet decomposition

Wavelet transformation has been widely used for analyzing signals. Its time-frequency localization (Daubechies, 1990) and multiresolution capability (Mallat, 1989) have proved extremely useful for applications of denoising (Donoho, 1995; Donoho and Johnstone, 1998; Donoho et al., 1995; Johnstone and Silverman, 1997), density estimation (Donoho et al., 1996; Kerkvacharian and Picard, 1992), data compression (Chambolle et al., 1998; Chang et al., 2000; Villasenor et al., 1995), and so on. More recently, it has been applied to genetic association analyses in high-throughput sequencing assays (Shim et al., 2015). In ribosome footprint profiles, there may be noise in sequencing and alignment, as well as insufficient sampling due to low sequencing depth. To extract reliable signals for downstream analysis, we adopt wavelet decomposition to build a multiresolution reconstruction of ribosome marginal densities.

Let \( p_i(x) \) denote the marginal density of ribosomes along a given transcript \( t \), such that \( p_i(x) \) represents the probability that the A-site of a randomly sampled ribosome is at position \( x \) (in units of codons). We can decompose this function as

\[
p_i(x) = \sum_k d_{i,k} \phi_{i,k}(x) + \sum_{i=2}^{\infty} \sum_k b_{i,k} \psi_{i,k}(x),
\]

where \( \phi_{m,n}(x) = 2^m/\sqrt{2^n} \psi(2^m x - n) \) and \( \psi_{m,n}(x) = 2^{m/2} \psi(2^m x - n) \), with \( \psi(x) \) and \( \phi(x) \) corresponding to a scaling function and its associated wavelet, respectively (Burrus et al., 1998). The coefficients \( d_{i,k} \) and \( b_{i,k} \) are uniquely determined by inner products \( \langle p_i(x), \phi_{i,k}(x) \rangle \) and \( \langle p_i(x), \psi_{i,k}(x) \rangle \), respectively. The spans of \( \phi_{m,n} \) and \( \psi_{m,n} \) form a multiresolution approximation of \( L^2(\mathbb{R}) \) as

\[
\{0\} \subset \ldots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset L^2(\mathbb{R}),
\]

where \( V_m = \text{span}(\{\psi_{m,n}\}_{n \in \mathbb{Z}}) \), \( W_m = \text{span}(\{\phi_{m,n}\}_{n \in \mathbb{Z}}) \), and \( V_m \oplus W_m = V_{m+1} \). Figure 2 shows the use of wavelet analysis to build a multiresolution reconstruction, in which global shapes are decoupled from local patterns. In what follows, we use \( p_i^h \) to denote the projection of \( p_i \) onto subspace \( S \). Table 1 summarizes the reproducibility of
Table 1. Correlation between the replicates projected onto various subspaces. The Pearson correlation coefficient between the replicates in GSE13750 increases as the scale (index $m$ of $V_m$) decreases.

<table>
<thead>
<tr>
<th>Length (codons)</th>
<th>Subspace $S$</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
<th>$V_5$</th>
<th>$V_6$</th>
<th>$V_7$</th>
<th>$L^2$($\mathbb{R}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–250</td>
<td>0.53</td>
<td>0.56</td>
<td>0.49</td>
<td>0.42</td>
<td>0.37</td>
<td>0.34</td>
<td>0.31</td>
<td>0.28</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>251–500</td>
<td>0.59</td>
<td>0.55</td>
<td>0.46</td>
<td>0.38</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.19</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>501–750</td>
<td>0.49</td>
<td>0.46</td>
<td>0.39</td>
<td>0.32</td>
<td>0.25</td>
<td>0.21</td>
<td>0.18</td>
<td>0.15</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>751–1000</td>
<td>0.51</td>
<td>0.44</td>
<td>0.34</td>
<td>0.27</td>
<td>0.21</td>
<td>0.18</td>
<td>0.15</td>
<td>0.12</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>1001–3745</td>
<td>0.53</td>
<td>0.45</td>
<td>0.36</td>
<td>0.27</td>
<td>0.21</td>
<td>0.17</td>
<td>0.14</td>
<td>0.12</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>

Note that, for each position $x \in \{1, \ldots, \ell\}$, there is exactly one codon type $j \in C$ such that $c_j^t[x] = 1$; all other strings will have a 0 at position $x$.

Next, we apply kernel smoothing to each string as

$$c_j^t[x] = \frac{\sum_{x'} K(x, x') c_j^t[x']} {\sum_{x'} K(x, x')}.$$  (2)

where $K(x, x')$ is a suitably chosen kernel. Intuitively, if the ribosome marginal density is affected by sequence context, then a region with a cluster of slow translating codons is likely to become congested with ribosomes. By smoothing $c_j^t$ with a kernel, we incorporate the distribution of neighboring codons into our model. Figure 3 illustrates the method we propose, in which the black spikes represent the decomposed codon sequences and the red curves are the kernel smoothed estimates with a symmetric Gaussian kernel, $k(x, x_t) = f_N(x, x_t, \sigma)$, where $f_N(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Since the flow of ribosomes is directional, a cluster of slow translating codons should have a larger impact on upstream ribosomes than on downstream ribosomes. To take this directionality into account, we therefore employ an asymmetric kernel. An example of asymmetric kernels is the probability density function of the gamma distribution,

$$f_C(x; \alpha, \beta) = x^{\alpha-1} e^{-x/\beta} / \Gamma(\alpha) \beta^{\alpha}, \quad x \in (0, \infty),$$

where $\alpha$ and $\beta$ are the shape and scale parameters, respectively, and $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the gamma function. It can be shown that the peak of this density function is located at $\theta = (\alpha - 1)\beta$, and the standard deviation is $\sigma = \beta \sqrt{\alpha}$. We define a modified gamma distribution $f_{BG}$, parameterized by $\theta$ and $\sigma$, by applying a horizontal shift.
and the modified gamma distribution reflecting so that the peak of the function is located at 0 and by reflecting about the y-axis:

\[ f_{\text{SRG}}(x; \theta, \sigma) = \frac{(-x + \theta)^{\alpha-1} e^{-(x - \theta)/\sigma}}{\Gamma(\alpha)\sigma^\alpha}, \]

where \( \alpha = 1 + \rho^2 - \sqrt{4\rho^2 + 2} \) and \( \theta = \frac{\rho}{\sqrt{4\rho^2 + 2}} \). The Gamma distribution \( f_\Gamma \) and the modified gamma distribution \( f_{\text{SRG}} \) are illustrated in Figure 4. We also define a shifted but non-reflected function \( f_{\text{SG}} \) to take into account the effect of positively charged residues of nascent peptides on the speed of downstream ribosomes (Charneski and Hurst, 2013):

\[ f_{\text{SG}}(x; \theta, \sigma) = \frac{(-x + \theta)^{\alpha-1} e^{-(x - \theta)/\sigma}}{\Gamma(\alpha)\sigma^\alpha}. \]

In what follows, we set the \( \theta \) parameter of \( f_{\text{SG}} \) and \( f_{\text{SRG}} \) to 5, because a ribosome occupies 9–10 codons and there are about 5 codons downstream. Applications of these asymmetric kernels \( f_{\text{SRG}} \) and \( f_{\text{SG}} \) are depicted as blue and green curves in Figure 3 respectively.

### 2.4 Sparse regression model

We relate the ribosome marginal density to the transcript sequence context by building a sparse model. In (2), we apply \( K(x, x') = f_{\text{SG}}(x - x', 0, \sigma_0) \) or \( K(x, x') = f_\Gamma(x, x', 0, \sigma_0) \) with \( B \) different bandwidths \( \sigma_1, \ldots, \sigma_B \). For each codon type \( j \in \mathbb{C} \), let \( \tilde{\beta} \) denote the kernel feature string smoothed using the kernel function \( f \in \{ f_\Gamma, f_{\text{SG}}, f_{\text{SRG}} \} \) with bandwidth \( \sigma \). If we use \( Q \) distinct kernel functions \( f_1, \ldots, f_Q \), then for each position \( x \) of transcript \( t \), the predictors can be represented as

\[ x_t(x) = (\tilde{c}_{1,0}(x), \tilde{c}_{1,1}(x), \ldots, \tilde{c}_{1,b}(x), \ldots), \]

\[ (\tilde{c}_{Q,0}(x), \tilde{c}_{Q,1}(x), \ldots, \tilde{c}_{Q,b}(x)) \in \mathbb{R}^{64 \times B \times Q}, \]

where \( \tilde{c}_{j,b}(x) = (\tilde{c}_{j,0,b}(x), \tilde{c}_{j,1,b}(x), \ldots, \tilde{c}_{j,63,b}(x)) \in \mathbb{R}^{64} \).

We relate these sequence features to the ribosome marginal density projected onto subspace \( \mathcal{S} \), by formulating a sparse regression problem:

\[
\min_\beta \sum_{t,x} \left[ p_t^a(x) - \frac{x_t(x)}{\ell_t} \right] k_{ij} + \lambda \| \beta \|_1, \tag{3}
\]

where \( \beta = (\beta_1, \ldots, \beta_{64 \times B \times Q}) \) denote regression coefficients. The first part \( \sum_{t,x} \left[ p_t^a(x) - \frac{x_t(x)}{\ell_t} \right]^2 \) of the objective function represents the residual sum of squares, whereas the second part \( \lambda \| \beta \|_1 \) promotes a sparse solution. The sparsity of the solution depends on the value of the regularization parameter \( \lambda \). This type of linear estimation, known as LASSO (Tibshirani, 1996; Zhao and Yu, 2006), avoids overfitting the data especially in scenarios where the number of predictors is much larger than the number of observations. It also provides an interpretation of the importance of each predictor, i.e., the variables that correspond to zeros in the solution \( \beta \) are not significant to the model. Our goal is to find a significant bandwidth \( \sigma_0 \). By training the sparse regression model with kernel smoothed codon feature strings \( \tilde{c}_{j,\sigma_0} \) with various bandwidths, and selecting the regularization parameter \( \lambda \) with cross validation to minimize mean squared error (MSE), the nonzero terms in the regression coefficients...
provide insight into the extent to which sequence context affects the ribosomal marginal density.

3 Results

After the filtering step described in Section 2.1, we grouped genes according to their length into 5 bins: 1–250 codons, 251–500 codons, 501–750 codons, 751–1000 codons and ≥1000 codons. We then decomposed each ribosome footprint profile using wavelet analysis with Daubechies-8 basis and eight scales, as illustrated in Figure 2. We considered a model with both asymmetric kernels \( f_{\text{AS}} \) and \( f_{\text{RSG}} \) each with eight bandwidths \( \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\} = (1, 3, 5, 7, 9, 11, 13, 15, 17, 19) \). For comparison, we also considered a model with the symmetric kernel \( f_s \) with the same set of bandwidths, as well as a model with no kernel smoothing. We trained the regression model using 9n/10 randomly chosen genes, where \( n \) is the number of genes in the corresponding bin, and chose the regularization parameter in (3) via five-fold cross-validation using the training data. Then the model was tested on the genes that were left out, and the process was repeated until every gene has been tested.

Results of CHX-treated footprints in various spaces are summarized in Table 2. The average performance is presented in terms of the Pearson correlation coefficient between the predicted and the true ribosomal marginal densities. Notice that our method using asymmetric kernels achieved high accuracy in subspaces \( V_6 \) through \( V_4 \). This suggests that the global shape of transcript-specific ribosome footprint profile is well captured by our proposed kernel-smoothed codon sequences. The performance in the subspace \( V_3 \) decreases as \( k \) increases. This is partly due to noise in the read alignment or the sequencing noise inherent in ribosome profiling. Thus, wavelet analysis is useful for studying the limit of the ribosome profiling technique and focusing on the reliable signals. The results from using non-smoothed codon sequences were the least accurate, supporting that the neighboring codons or interference play an important role in ribosome footprint distributions. Using asymmetric kernel smoothing produced more accurate results than using a symmetric kernel. The same trend also holds for flash-freeze ribosome profiles (Table 2). The performance of our method on CHX-treated ribosome profiles was better than that on the flash-freeze profiles. This may be due to the large amplitude of ramps near the 5' end in CHX-treated profiles (Weinberg et al., 2016). These ramps contribute significantly to global shapes, and can be well captured by asymmetric kernel smoothing.

In general, our method with asymmetric kernels was able to predict the global shape of transcript-specific ribosome density with high accuracy. However, for the flash-freeze data, the correlation in the subspace \( V_3 \) was relatively poor for the genes shorter than 250 codons. This is because the global shape of their profiles in this subspace was rather flat for the flash-freeze data. If the mean squared error is used as the performance measure, the results for shorter genes are not so much worse than that for longer genes. As an example, Figure 5 shows the ribosome density of YBR212W in various subspaces and the corresponding predictions for both CHX-treated and flash-freeze profiles. Figure 6 shows heatmaps in various subspaces and the corresponding predictions for both CHX-treated and flash-freeze profiles. This may be related to wobble base pairing, which has been found to be important in the ribosome profile shape of YBR212W (Stadler et al., 2014). Furthermore, CGA (arginine) and CGG (arginine) have a strong influence on the global shape of ribosome occupancy in flash-freeze data, while CCA and CGG seem to have the dominant effect in CHX-treated data. This may be related to wobble base pairing, which has been found to be associated with elongation (Lareau et al., 2014; Stadler and Fire, 2011). In particular, high ribosome occupancy has been observed on wobble-paired proline CCG (G-U base pairing) and arginine CCA (I-A pairing) by Lareau et al. (2014). Furthermore, CCA is one of the codons with the most prolonged elongation rates, consistent with Lareau et al.

To examine further how the distribution of codons may affect the shape of ribosome footprint profiles, we carried out a sliding-window analysis for each transcript, using a window size of 20 codons. For each transcript, we first identified all windows for which the average ribosome density was larger than twice the transcript-wide average ribosome density, and then considered those windows for codon enrichment analysis. For each selected window, let \( \omega = \{o_1, o_2, \ldots, o_{40}\} \) denote the observed codon distribution, where \( o_i \) is the number of codons of type \( i \) within the window. We compared this observed codon distribution with the expected
distribution \( \eta = [\eta_1, \eta_2, \ldots, \eta_{64}] \), where \( \eta_i \) denotes the expected number of codons of type \( i \) within the window, assuming sampling 20 codons without replacement according to the codon distribution in the entire transcript. Note that \( \eta \) follows a multivariate hypergeometric distribution.

We performed a paired \( t \)-test on \( \omega \) and \( \eta \) over the selected windows and identified the codons enriched in high ribosome-occupancy regions. We then compared the \( P \)-values from the enrichment analysis with the frequencies of the regression coefficients being nonzero, as illustrated in Figure 7. The scatter plot for flash-freeze data (Fig. 7B) shows that all of the significantly enriched codons were selected frequently by the regression model. Codons that were selected frequently but not significantly enriched in high density regions might have been selected for prediction in low ribosomal density regions. For example, it has been observed previously (Lareau et al., 2014) that GGG and GTT are some of the codons with the least footprint abundance. For the CHX-treated data, there were a few codons enriched significantly but not selected frequently (Fig. 7A); this could be due to technical biases in the data, discussed in Section 4.

4 Discussion

Our results on \textit{S. cerevisiae} data, especially the CHX-treated footprints, show that the global shape of transcript-specific ribosome
density can be predicted with high accuracy. This suggests that the ribosome distribution along a transcript indeed depends on the sequence context. In particular, we find a few codons that dominate the global shape; these codons influence the ribosome density not only at the site where they appear, but also their neighboring positions.

There are several possible reasons why applying kernel smoothing to codon sequences helps in predicting the shape of transcript-specific ribosome occupancy profile. One possible reason is the interference among ribosomes. Another possible explanation is that the physical footprint of each ribosome spans multiple codons (9–10 codons). Lastly, it has been shown (Charneski and Hurst, 2013) that positively charged amino acids on nascent peptides significantly slow down ribosomes downstream from where the residues are encoded. Among all the methods compared, the asymmetric kernel smoothing performed the best. This could be due to the directional movement of ribosomes and the aforementioned effect of positively charged amino acids on downstream elongation speeds. Also, 5’ ramps can be better predicted using asymmetric kernels, which partly explains why the prediction accuracy for CHX-treated ribosome profiles was slightly better than that for flash-freeze profiles.

There are other factors not incorporated into our current model that could also help to explain ribosome distributions, e.g. mRNA secondary structure. Notice that in the results for CHX-treated ribosome ramps can be better predicted using asymmetric kernels, which partly explains why the prediction accuracy for CHX-treated ribosome distributions is lower near the 5’ end. Furthermore, when computing differences between the observed ribosome densities and the predicted densities from these poorly predicted genes, we noticed a weak positive correlation (r = 0.0446) between the differences and the likelihood of double stranded conformations, i.e., the PARS (parallel analysis of RNA structure) scores of secondary structure defined by Kertesz et al. (2010) in a genome-wide measurement. Although many potential factors— including mRNA secondary structure, codon usage bias, tRNA- adaption index, etc.—are not explicitly incorporated into our model, most of these factors depend on the sequence context. This may be one of the reasons why considering transcript sequences alone enables reasonably accurate prediction of ribosome footprint profile shapes.

It is possible that some of the features learned by our predictive model are due to technical biases in the data. Library construction is a possible source of bias due to technical artifacts. Artieri and Fraser (2014) found three types of nucleotide biases by comparing mRNA fragments (unprotected by ribosomes) with ribosome footprints to identify shared biases. These include an enrichment of adenine at the 3’ and 5’ ends, and a depletion of cytosine at the fourth nucleotide position in both the mRNA and ribosome footprint reads. However, we note that the flash-freeze ribosome footprint data (Weinberg et al., 2016) we analyzed were generated using a comprehensive set of adapter sequences to avoid the biases discussed in Artieri and Fraser (2014). Furthermore, the flash-freeze protocol avoids the ribosome run-off phenomenon that can occur in CHX pre-treatment.

Different models have been proposed in the literature to estimate elongation rates. Shah et al. (2013) developed a continuous-time, discrete-state Markov model to simulate translational dynamics, in which initiation and elongation rates were inferred using the cell volume, ribosome abundance and tRNA abundance. These elongation rates were codon dependent but not position dependent. The model proposed by Pop et al. (2014) was based on flow conservation, and it assumed that the occurrences of the same codon within an mRNA transcript have the same dwell time, and hence the same elongation rate. Our results suggest that these models should take the codon position into account. In other words, the elongation rate at a particular position along a transcript not only depends on the codon at that position but also the neighboring codons and how the neighboring codons are ordered.
Part of the results presented in this paper is based on combining four publicly available datasets. This overcomes the difficulty of sparse sampling that other work may encounter, e.g., Pop et al. (2014). Although our study is limited to the steady-state distribution of ribosomes, it examines the extent to which the sequence context influences ribosome distributions under normal conditions. One of the important applications of our method is to examine the positions where the steady-state prediction is significantly different from the actual measurements. These positions may lead to important findings of unknown factors that regulate translation. For example, our proposed method can be applied to cells grown under various environmental conditions, or to ribosome profiles obtained by different inhibitors, e.g. harringtonine (Ingolia et al., 2011), anisomycin (Lareau et al., 2014).

Another application is the inference of isoform-specific ribosome footprints. It is known that many human genes have more than one isoform, and these isoforms contribute to the complexity of phenotypes (Johnson et al., 2003; Lander et al., 2001; Wang et al., 2008). Studying the translation of specific isoforms requires correctly assigning ribosome footprints to them. However, since the footprints of ribosomes are short (28–30 nucleotides), identifying the correct isoform becomes challenging. Our method provides a likelihood function for ribosome location, and can facilitate sequence alignment when more than one alignment exists. Synthetic biology is another active research area (Endy, 2005) that can benefit from our work: the ability to predict and compare the ribosome densities along different transcripts under various environmental conditions can facilitate the search for sequences with optimal translational properties and reduce the number of experimental trials necessary to build and test the sequences.

**Funding**

This research is supported in part by an NSF CAREER Grant DBI-0846015, a Packard Fellowship for Science and Engineering and a MathX Research Grant from the Simons Foundation.

**Conflict of Interest:** none declared.

**References**


Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/12/i183/2240607 by guest on 08 January 2019

