Scikit-ribo - Accurate A-site prediction and robust modeling of translational control

Han Fang

February 12, 2016
AGBT
Acknowledgments

**Lyon Lab**
Max Doerfel  
Yiyang Wu  
Jonathan Crain  
Jason O’Rawe

**Schatz Lab**
Fritz Sedlazeck  
Tyler Garvin  
James Gurtowski  
Maria Nattestad  
Srividya Ramakrishnan

**Cold Spring Harbor Laboratory:**
Yifei Huang  
Noah Dukler  
Adam Siepel  
Melissa Kramer

**Stony Brook University:**
Eric Antoniou  
Elena Ghiban  
Stephanie Muller  
Rob Patro
Central dogma of biology – Classic view

DNA \rightarrow RNA \rightarrow Protein

Replication: DNA → RNA

Transcription: RNA Polymerase

Translation: Ribosome
What is ribosome profiling (Riboseq)?

Calculate translational efficiency (TE)

**Less** efficient translation

![Less efficient translation diagram](image)

\[ \log_2(TE) < 0 \]

**Normal** translation efficiency (TE)

![Normal efficiency diagram](image)

\[ \log_2(TE) = 0 \]

**More** efficient translation

![More efficient translation diagram](image)

\[ \log_2(TE) > 0 \]

\[
TE = \frac{\text{Riboseq rpkms}}{\text{RNAseq rpkms}}
\]
Hypothesis: TE distribution could be skewed by ribosome pausing events.

Ribosome footprints without bias

Ribosome footprints with pausing
Simulated *S. cerevisiae* data - TE distribution are negatively-skewed by ribosome pausing events

\[
\text{TE} = \frac{\text{Riboseq rpkm}}{\text{RNAseq rpkm}}
\]
Analytical Challenges

- Understand translational control
- Assay specific characteristics/biases (e.g., ribosome pausing)
- Actively translated codons

- How to accurately infer translation efficiency?
- How is Riboseq different from RNAseq?
- Where does the A-site locate on Riboseq reads?
Introducing scikit-ribo

- Ribosome A-site position prediction
- A-site codon localization
- Ribosome pausing site calling
- Translation efficiency inference
- Differential translation efficiency testing
What and where is the ribosome A-site?

Figure adapted from Ingolia et al. Science (2009)
How to predict A-site?

Training data and features:

Classifier and model tuning:

• SVM with RBF kernel (scikit-learn)
• 10 fold cross-validation for grid search
• Make predictions on all reads genome-wide
Scikit-ribo has much higher accuracy of identifying A-site than the previous method (0.86 vs. 0.64, 10-fold CV).
Scikit-ribo accurately predicted codon usage fraction and codon normalized TE.

Finding ribosome pausing sites (peaks) is hard. But it is easier after knowing the A-site location.

Q: how to robustly identify ribosome pausing sites while accounting for over-dispersion?
Ribosome pausing site identification by negative binomial mixture model

\[ P(X_i|\pi_i, \mu_i, k_i, r_i) = \prod_j \pi_i NB(X_{ij}|\mu_i, r_i) + (1 - \pi_i) NB(X_{ij}|k_i\mu_i, r_i), \]

for gene i at position j, where \( k \geq 5 \)

\( H_0 : \pi = 1 \)
\( H_1 : \pi \neq 1 \)
Ribosome pausing site identification by negative binomial mixture model

\[ P(X_i | \pi_i, \mu_i, k_i, r_i) = \prod_j \pi_i N B(X_{ij} | \mu_i, r_i) + (1 - \pi_i) N B(X_{ij} | k_i \mu_i, r_i), \]

for gene i at position j, where \( k \geq 5 \)

\( H_0 : \ \pi = 1 \)
\( H_1 : \ \pi \neq 1 \)

<table>
<thead>
<tr>
<th># genes</th>
<th># genes (rpkm &gt; 100)</th>
<th># genes with pausing</th>
<th># ribosome pausing sites identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>6664</td>
<td>1252</td>
<td>94</td>
<td>180</td>
</tr>
</tbody>
</table>
mRNA with stronger secondary structure tend to have ribosome pausing events

PARS scores obtained from Kertesz et al. Nature (2010)
TE distributions are negatively-skewed in many studies. Over-structured mRNA show inflated TE.

Chi Square test p-value < \(2 \times 10^{-16}\)

Weinberg, Shah et al. (2015)
Summary

Discussed:
1) Introduce scikit-ribo for joint analysis of Riboseq & RNAseq data.
2) Learn from data itself to determine ribosome A-site location.
3) Reveal biases in Riboseq data due to ribosome pausing.
4) How Riboseq biases lead to issues with estimating TE.

Ongoing work:
1) Joint inference of codon elongation rates and protein TE.
2) Extend the ribosome pausing calling to a HMM based method.

https://github.com/hanfang/