Model Selection for Bias Correction in RNA-Seq

Adam Roberts Harold Pimentel Matthias Vallentin

{adarob,pimentel,vallentin}@cs.berkeley.edu

December 13, 2010

The biochemistry of fragmentation, amplification and size selection in RNA-Seq experiments has been shown to lead to the over- or under-representation of fragments containing certain sequence motifs and to deviations from the expected uniformity of fragments across transcripts. These biases affect fragment counts and must be accounted for when estimating expression levels of individual transcripts.

We apply model selection techniques to this task of modeling the bias, and use a method similar to importance sampling to remove it. We find improvements in expression estimates as measured by correlation with independently performed qRT-PCR, and compare the models found using different criterion (AIC, BIC, likelihood) with the independent sites model.

1 Introduction and Background

RNA-Seq technology offers the possibility of accurately measuring transcript abundances in a sample of RNA by sequencing of double stranded cDNA [8]. Unfortunately, current technological limitations of sequencers require that the cDNA molecules represent only partial fragments of the RNA being probed. The cDNA fragments are obtained by a series of steps, including reverse transcription primed by random hexamers (RH), or by oligo(dT). Most protocols also include a fragmentation step, typically RNA hydrolysis or nebulization, or alternatively cDNA fragmentation by DNase I treatment or sonication. Many sequencing technologies also require constrained cDNA lengths, so a final gel cutting step for size selection may be included. Figure 1 shows how some of these procedures are combined in a typical experiment.

The randomness inherent in many of the preparation steps for RNA-Seq leads to fragments whose starting points (relative to the transcripts from which they were sequenced) appear to be chosen *approximately* uniformly at random. This observation has been the basis of assumptions underlying a number of RNA-Seq analysis approaches that invert the “reduction” of transcriptome estimation to DNA sequencing [5, 6, 9, 10, 14]. However, recent carefully analysis has revealed sequence-specific [4, 13] biases in sequenced fragments. Sequence-specific bias is a

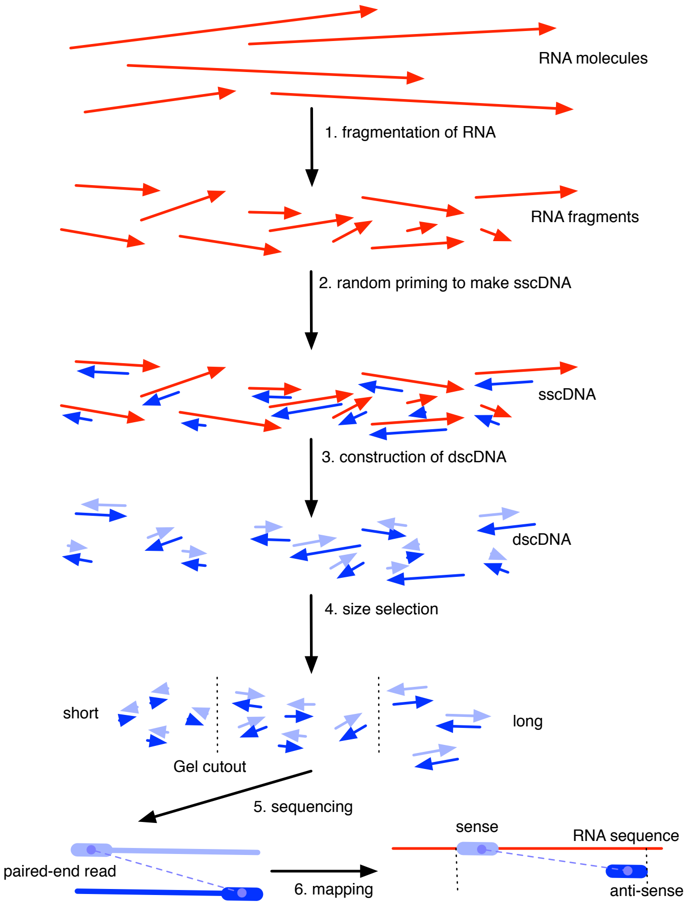


Figure 1: Overview of a typical RNA-Seq experiment. RNA is initially fragmented (1) followed by priming to make single stranded cDNA (2). Double stranded cDNA (3) is then size selected (4) resulting in fragments suitable for sequencing (5). Sequenced reads are mapped to the genome (6), and in the case of known transcript or fragment strandedness, the read alignments reveal the $5'$ and $3'$ ends of the sequenced fragment (see Supplementary Material Section 7.2). All arrows are oriented $5'$ to $3'$.

global effect where the sequence surrounding the beginning or end of potential fragments affects their likelihood of being selected for sequencing, often do to the use of random hexamer primers which have different priming affinities. An example of such bias can be seen in Figure 2. These biases can affect expression estimates [7], and it is therefore important to correct for them during RNA-Seq analysis.

Although many biases can be traced back to specifics of the preparation protocols, it is currently not possible to predict fragment distributions directly from a protocol. This is due to many factors, including uncertainty in the biochemistry of many steps and the unknown shape and effect of RNA secondary structure on certain procedures [7]. It is therefore desirable to estimate the extent and nature of bias indirectly by inferring it from the data (fragment alignments) in an experiment. However, such inference is non-trivial due to the fact that fragment abundances are

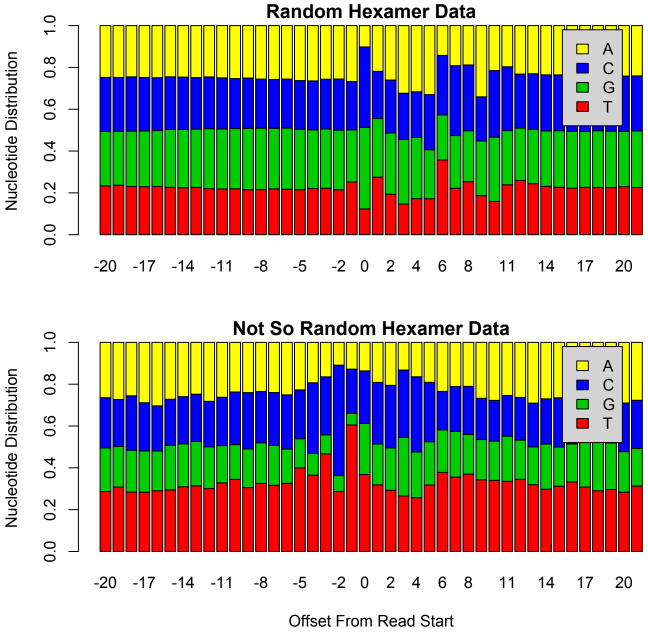


Figure 2: The nucleotide frequencies surrounding the fragment end for an experiment primed with random hexamers (top) [2] and 'not so random hexamers' (bottom) [1]. Note that most of the bias lies within a 21bp window extending from 8bp upstream and 12 bp downstream of the start site (0).

proportional to transcript abundances, so that the expression levels of transcripts from which fragments originate must be taken into account when estimating bias. At the same time, expression estimates made without correcting for bias may lead to the over or under-representation of fragments. Therefore the problems of bias estimation and expression estimation are fundamentally linked, and must be solved together. Likelihood based approaches are well suited to resolving this difficulty, as the bias and abundance parameters can be estimated jointly by maximizing a likelihood function for the data.

In [11], we developed a likelihood based approach for simultaneous estimation of bias parameters and expression levels using the likelihood framework of [14]. However, we now add a model selection step in which we search for a model of the sequence motif that takes into account correlations among nucleotide positions in the read and maximizes the likelihood. We demonstrate that our method improves expression estimates over the uniform model of [14] as well as a simple bias model where positions are assumed independent [11], in comparison with qRT-PCR on a benchmark dataset. Using the same data, we also show that our method improves on the approach of [4] where overrepresentation and underrepresentation of fragment ends was essentially measured using a fully connected graph to model the first 7 bases of every read.

2 Model Selection

2.1 Abundance Likelihood

The likelihood model used for abundance estimation extends the model in [14] and descried fully in [11], which is mainly outside of the scope of this paper. We employ an approach similar to importance sampling to weight the original model, which assumes uniform distribution of reads, by the actual distribution based on the sequences surrounding the ends of reads. Thus we require a model for these sequences that will allow us to tractably estimate this distribution.

2.2 Model Likelihood

In order to make the model selection step tractable, we make several reasonable simplifying assumptions. First, we replace the generative model of [14] used above with one where fragments are selected for priming directly from the pool of transcripts without previously selecting a gene, transcript, and fragment length. Second, we assume the two ends of the fragments are chosen for priming independently. Thus we can model each end independently and take the product of their probabilities to weight the uniform distribution by. Third, we assume there are only local correlations between positions. In the case of bias caused by the use random hexamer primers, this assumption is a good one, as the binding affinities of sliding windows of 6bp sequences will explain most of the bias. This assumption allows us to look for correlations in small windows instead of over the entire bias window. Finally, we also assume that the model can be represented by a DAG. While there is not much biological grounding for this assumption, it is required to make the solution tractable, and it should still capture much of the correlations between positions.

We define a DAG model $M = (X, \pi, \theta)$ for the $5'$ end (w.l.o.g.) where the nodes representing nucleotide positions in the window surrounding the read start are specified in $X = X_1, ..., X_N$, the parents are specified in π, and the conditional probabilities are the parameters $\theta_{X_i, X_{\pi_i}} = p(X_i \,|\, X_{\pi_i})$ Given the assumptions above, we write the likelihood of M given the observed $5'$ fragment ends (F) as

$$\begin{aligned}
L(M \mid F) &= \prod_{f \in F} p(f \mid M) \\
&= \prod_{f \in F} \prod_{i=1}^{N} p(X_i = f_i) \\
&= \prod_{f \in F} \prod_{i=1}^{N} p(X_i = f_i \mid X_{\pi_i} = f_{\pi_i}) \\
&= \prod_{i=1}^{N} \prod_{f \in F} p(X_i = f_i \mid X_{\pi_i} = f_{\pi_i}) \\
&= \prod_{i=1}^{N} \prod_{x_i, x_{\pi_i}} p(x_i \mid x_{\pi_i})^{m(x_i, x_{\pi_i})} \\
&= \prod_{i=1}^{N} \prod_{x_i, x_{\pi_i}} \theta_{x_i, x_{\pi_i}}^{m(x_i, x_{\pi_i})}
\end{aligned}$$

where $m(x_i, x_{\pi_i}) = \sum_{f \in F} \delta(f_i, x_i)\delta(f_{\pi_i}, x_{\pi_i})$.

2.3 Likelihood Maximization

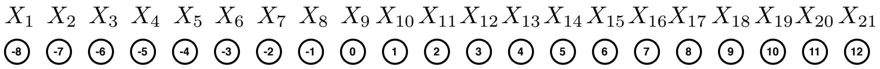

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14} X_{15} X_{16} X_{17} X_{18} X_{19} X_{20} X_{21}
-8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12

Figure 3: The independent sites model, which acts as a base for our model selection. The numbers in the nodes represent the offset from the read start (0).

For our model, we have set a fixed X as the 21bp window surrounding the $5'$ fragment end as shown in Figure 3. This decision was made based on our observations of the variability of nucleotide frequencies in this window (Figure 2). With X fixed, we wish to find the π, θ pair that maximizes the likelihood above. We apply coordinate ascent to search for a model that maximizes the likelihood above. Given fixed x_{π_i}, there are a discrete number of choices for X_i, namely the nucleotides A, C, T, or G. Therefore, $X_i \mid x_{\pi_i} \sim Multinomial(\theta_{x_i, x_{\pi_i}})$ which implies that the MLE is

$$\widehat{\theta}_{x_i, x_{\pi_i}} = \frac{m(x_i, x_{\pi_i})}{m(x_{\pi_i})}$$

The likelihood for a model M with fixed X is given by

$$L(\pi, \theta \,|\, X, F) = \prod_{i=1}^{N} \prod_{x_i, x_{\pi_i}} \theta_{x_i, x_{\pi_i}}^{m(x_i, x_{\pi_i})}$$

and the maximum likelihood for a specific $\bar{\pi}$ is

$$\widehat{L}(\bar{\pi} \,|\, X, F) = \prod_{i=1}^{N} \prod_{x_i, x_{\bar{\pi}_i}} \widehat{\theta}_{x_i, x_{\bar{\pi}_i}} = \prod_{i=1}^{N} \prod_{x_i, x_{\bar{\pi}_i}} \frac{m(x_i, x_{\bar{\pi}_i})}{m(x_{\bar{\pi}_i})}$$

Finally, this allows us to write the maximum log-likelihood for $\bar{\pi} \,|\, X, F$ as

$$\widehat{l}(\bar{\pi} \,|\, X, F) = \log\left\{\widehat{L}(\bar{\pi} \,|\, X, F)\right\} = \sum_{i=1}^{N} \sum_{x_i, x_{\bar{\pi}_i}} m(x_i, x_{\bar{\pi}_i}) \log\left\{\frac{m(x_i, x_{\bar{\pi}_i})}{m(x_{\bar{\pi}_i})}\right\}$$

2.4 Search Procedure

We wish to find $\widehat{\pi} = \arg\max_{\bar{\pi}} l(\bar{\pi} \,|\, X, F)$. We use *hill-climbing with restarts* to find the best local optimum out of many climbs using random starting graphs. At each step, we modify the current graph in a manner that increases the score most under the search criteria (log-likelihood, AIC, BIC). This modification can come from either removing or adding an edge while maintaining the acyclic property. Under the assumption of local correlations, we are also able to limit the search space by only adding edges less than a specified length, which we have set as three for our experiments.

3 Implementation

We implemented the graphical model, likelihood maximization, and hill climbing search procedure in 1589 lines of C++ code with extensive use of the Boost libraries [3]. Our program, called `outdoor`, provides fast data structures to *(i)* represent DAGs, *(ii)* compute likelihoods, and *(iii)* search the model space in a multi-threaded fashion. To ensure the correctness of implementation, we equipped `outdoor` with a unit test suite for the fundamental data structures.

The core of our implementation is an efficient data structure based on bit vectors that lies between an adjacency list and adjacency matrix. Because many high-level graph operations, such as adding edges or enumerating the children of a node, translate to logical bitwise manipulations we are able reap high performance gains. This data structure is geared towards a DAG that has a maximum distance of k to its neighbors where $k < N$. That is, given a total ordering on the nodes X_i, the neighbor set $\mathcal{N}(i)$ of node i must satisfy $\mathcal{N}(i) \subseteq \{\min\{0, i-k\}, \ldots, \max\{i+k, N\}\}$.

Consequently we represent each node with $2k$ bits to encode the edges to its neighbors, where the first $[0, k)$ bits are for backward edges and the remaining $[k, 2k)$ bits for forward edges.[1]

For a given graph, the likelihood calculation is handled by the *sherpa*, a wrapper who interfaces to the table containing the counts $m(x_i, x_{\bar{\pi}_i})$. This $N - 2k$ column table is a base-5 representation of the counts of a sequence containing all combinations of A, C, G, T, *, where * is a wildcard. The columns represent a sliding window of distance from the beginning of the sequence end. Thus, the memory space required for the table is $(N-2k)5^k$. The wildcard allows for sequences such as *GAC*A*A, where * can be any sequence, and thus is the sum of all possible combinations at that position. Our code handles the likelihood computation efficiently by precomputing the *entire* table. Once computed, the sherpa calculates the likelihood given an arbitrary graph and the table by simply doing a series of sums and constant time lookups.

The main purpose of `outdoor` is to search the model space according to a given metric by choosing a random point in the model space, climbing uphill to a local peak, and then with restarting (see §2.4). To this end, the *base camp* component organizes *expeditions*, i.e., launches multiple climbers in parallel, each of which are being sent on a journey to find a local peak given a random starting graph. Every climber runs as a separate thread and we found that setting the number of threads slightly higher than the available cores (+1/2) yields the best performance results. The base camp records the local peaks from each expedition and selects the best model amongst them as global peak. In the next section we discuss the shape of the global peaks we found in our data sets.

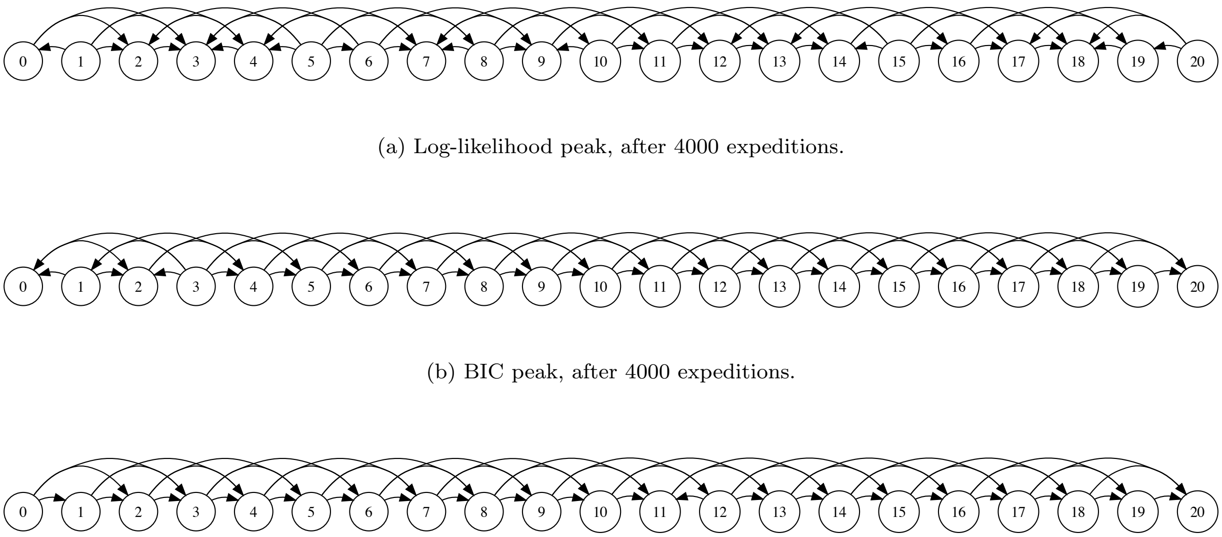
4 Results

4.1 Resulting Models

Using the likelihood of §2.2 and search procedure of §2.4, we ran our algorithm for several criteria (log-likelihood, AIC, and BIC) with 4000 random starts on the Illumina RNA-Seq dataset, which was sequenced from a standard Human Brain Reference (HBR) sample using random hexamers (RH) and was published in [2]. The nucleotide frequencies for this dataset can be found in Figure 2. In Figure 4 we present the best models found under each criterion, and in the top row of Figure 5 the path taken by the hill-climbing algorithm from the random start that found the peak. As expected, the log-likelihood criteria produces a model with the most edges, since there is no penalty for how complicated the model is. In this application, over-fitting may not be an issue since we will only apply the model to correction of the same dataset.

We also ran the same experiments on a different RNA-Seq dataset which sequenced the same HBR sample using

[1] Technically, the first k bits of the node 0, $k-1$ bits of node 1, etc., are not needed, but we still kept them in our initial prototype to simplify the implementation. This also applies to the k last nodes. Regardless, the waste of $k(k+1)$ bits is not problematic because our program is not memory-bound.



(a) Log-likelihood peak, after 4000 expeditions.

(b) BIC peak, after 4000 expeditions.

(c) AIC peak, after 4000 expeditions.

Figure 4: Peaks for the RH data set.

'not so random' primers (NRH) [1]. The nucleotide frequencies for this dataset can be found in Figure 2 and the selected models and paths in Figure 6. Note that the use of different primers has led to different models being selected when all other experimental variables are held constant.

The bottom rows of Figure 5 and Figure 7 show the empirical distribution of local peaks for the different metrics. Each discontinuity in the ECDF corresponds to a mode and means that several climbers have found a similar optimum. For the RH dataset in Figure 5, observe that the mode of the log-likelihood metric (left-most panel) is not located at the global optimum, i.e., the righ-most value in the ECDF. However, the AIC and BIC modes are exactly at the global optimum. For the NRH dataset in Figure 7, no mode is located at the global optimum.

4.2 Evaluation

To evaluate the resulting models, we applied them to the bias correction method mentioned in §2.1 and described fully in [11]. We found the abundance estimates for the two RNA-Seq methods mentioned above using the uniform model of [14], the independent sites bias correction model of [11], and the models produced by the log-likelihood, AIC, and BIC criteria within the framework of [11]. In Figure 8 we present the R^2 value from the correlation of these abundance estimates with qRT-PCR for 907 transcripts obtained from the Microarray Quality Control (MAQC) dataset [12], which used the same HBR sample.

The model selection obviously improves upon the independent model. We note that using any one of the three new methods improves the R^2 value. In the NRH data, AIC and BIC performed very similarly. Looking at the bar

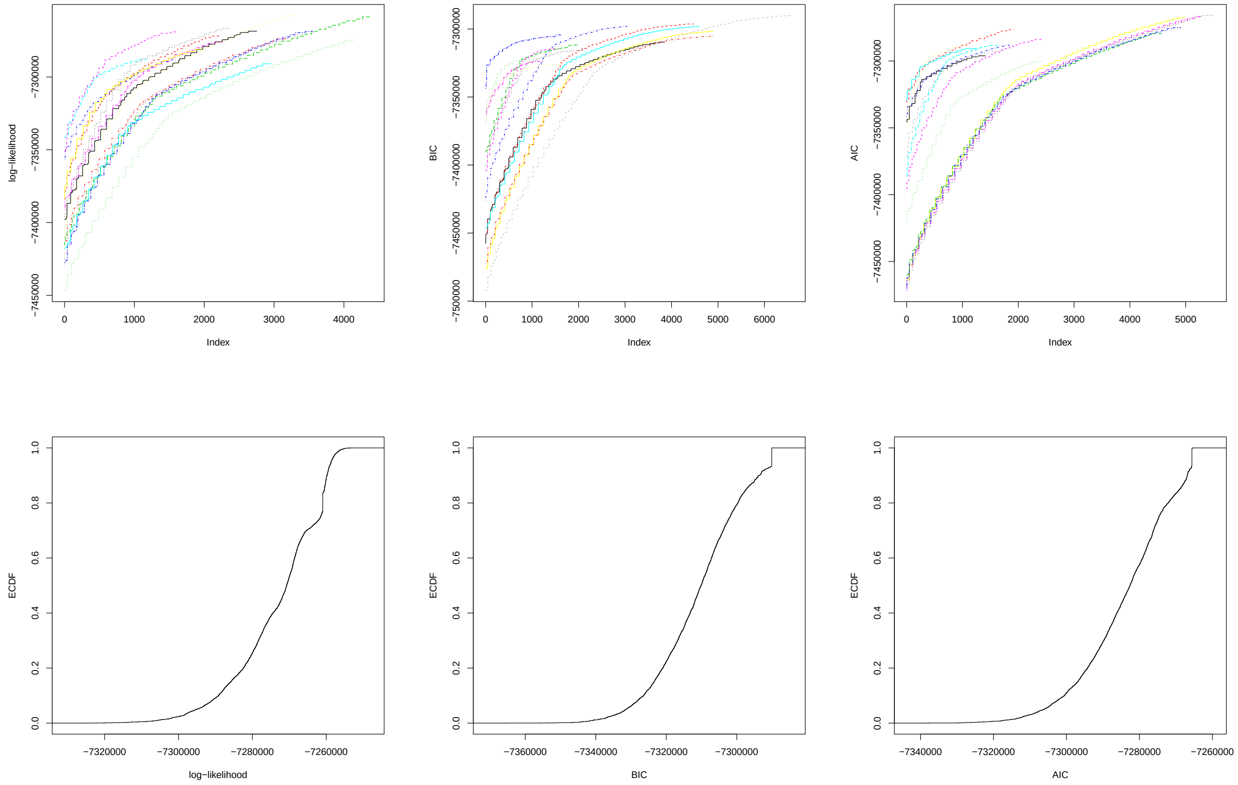
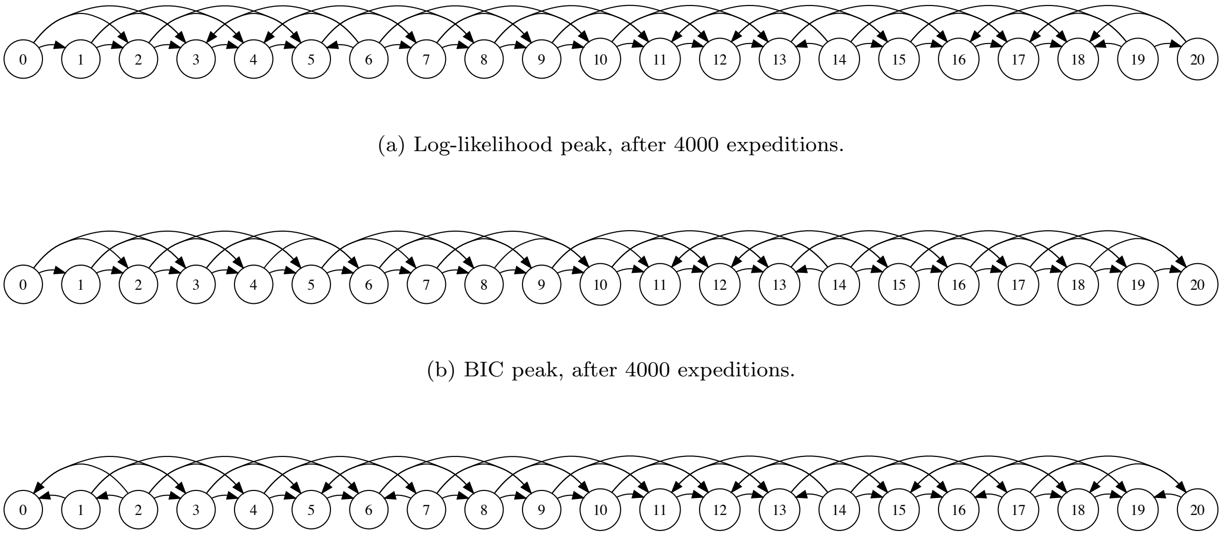


Figure 5: Hill climbing results of the RH data set. The first row shows 15 random climbers for three metrics log-likelihood, BIC, and AIC. The second row shows the distribution of local peaks.

graph we note that the resulting graphs are very similar with very few difference in edges. As expected, the model using only the log-likelihood is a very densely connected graph. The reason for this is that there is no penalty for adding additional edges and the likelihood is increased by adding additional terms.

We note that the resulting AIC and BIC models are very similar with the exception of direction of edges. A possible explanation is that interactions mostly act downstream, rather than upstream, as we see very few connections going back.

In the NRH dataset, we note the AIC and BIC models are again, very similar, and even similar to the results of the RH dataset. In this case, there are more backward edges implying that at certain interactions are stronger in the backward direction than forward. Finally, we mention that in these two cases, AIC appears to have the most consistently high R^2 of both datasets.



(a) Log-likelihood peak, after 4000 expeditions.

(b) BIC peak, after 4000 expeditions.

(c) AIC peak, after 4000 expeditions.

Figure 6: Peaks for the NRH data set.

5 Conclusion

The results of this experiment show that using model selection improves the bias correction of [11]. As different RNA-Seq experiments are likely to have different sequence biases based on the exact protocol used, it is necessary to select a new model for each new analysis. The additional model selection step appears to be promising as the new step results in graphs with higher R^2 than previous methods. Particularly, the AIC criterion appears to have the highest consistent R^2.

We find it interesting that in general, the resulting models have edges in the forward direction, and very few going back. Preliminarily, this leads us to believe that perhaps bias influence of nucleotides is downstream, rather than upstream.

There are several additional steps we would like to elaborate on in this project resulting from the current discoveries. Firstly, we would like to test our method on several additional datasets which have corresponding qRT-PCR data. Second, we would like to explore graph with edges with greater distance than k in each direction. A reason for this is that several of the edges in all the resulting models span the greatest possible distance of k in each direction. Our current constraints do not allow us to capture longer interactions, which, given our current results seem plausible. Finally, we would like to implement undirected model selection, as we would like to compare the undirected interactions to directed.

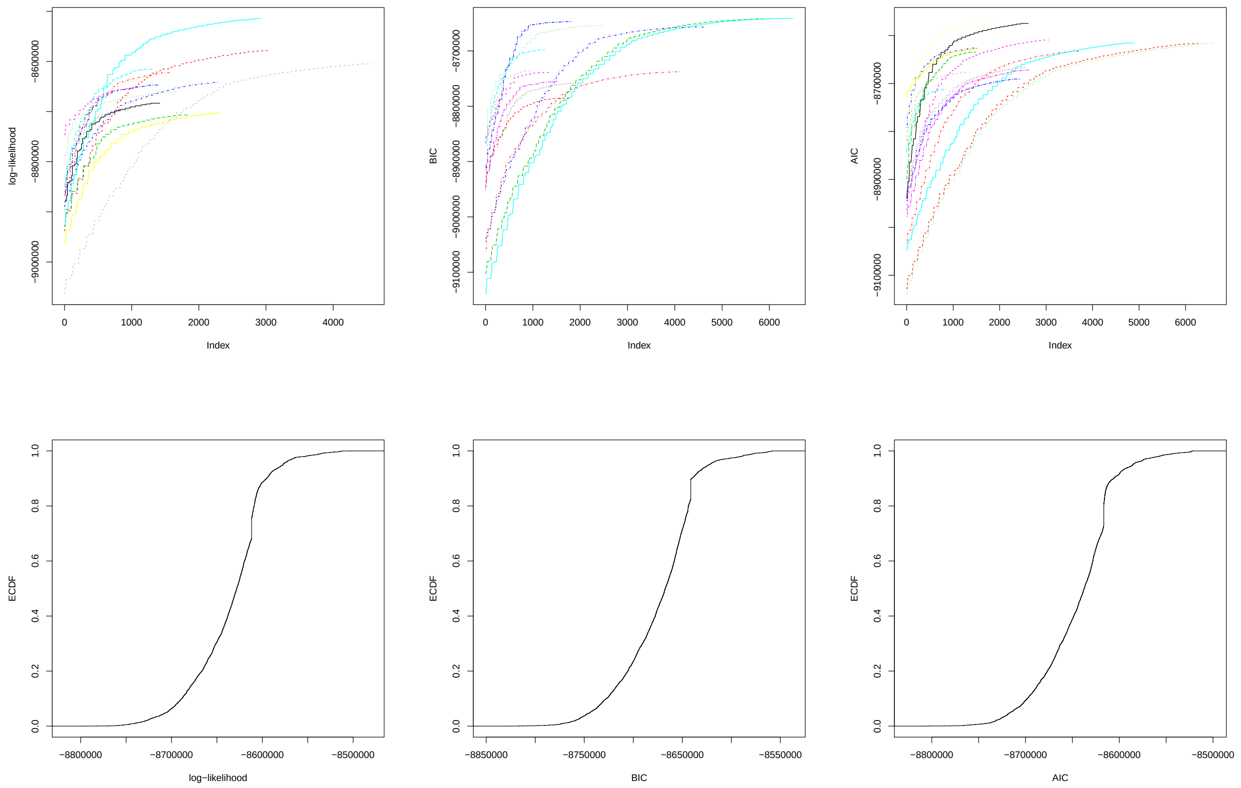


Figure 7: Hill climbing results of the NRH data set. The first row shows random number of climbers for three metrics log-likelihood, BIC, and AIC. The second row shows the distribution of local peaks.

Acknowledgements

We acknowledge Lior Pachter for helping devise the project.

References

[1] C. Armour, J. Castle, R. Chen, T. Babak, P. Loerch, S. Jackson, J. Shah, J. Dey, C. Rohl, J. Johnson, and C. Raymond. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods*, 6:647–649, 2009.

[2] K. Au, H. Jiang, L. Lin, Y. Xing, and W. Wong. Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. *Nucleic Acids Research*, 38:4570–4578, 2010.

[3] Boost C++ Libraries. `http://www.boost.org`.

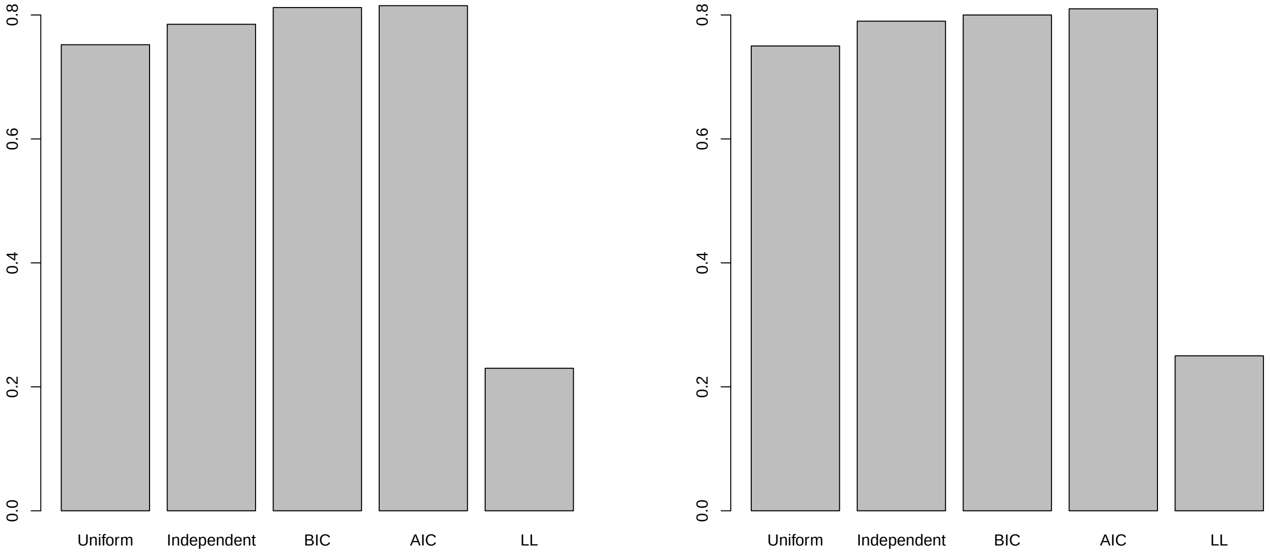


Figure 8: R^2 of model selection criterion vs qRT-PCR data. The left panel is RH data, and the right is NRH.

[4] K. Hansen, S. Brenner, and S. Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38:1–7, Apr 2010.

[5] H. Jiang and W. Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25:1026–1032, Jan 2009.

[6] B. Li, V. Ruotti, R. Stewart, J. Thomson, and C. Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26:493–500, Feb 2010.

[7] J. Li, H. Jiang, and W. Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol*, 11:R50, May 2010.

[8] S. Marguerat and J. Bähler. Rna-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 67:569–579, Jan 2010.

[9] M. Nicolae, S. Mangul, I. Măndoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms in Bioinformatics*, 6293:202–214, 2010.

[10] B. Paaniuc, N. Zaitlen, and E. Halperin. Accurate estimation of expression levels of homologous genes in rna-seq experiments. In B. Berger, editor, *Research in Computational Molecular Biology*, volume 6044 of *Lecture Notes in Computer Science*, pages 397–409. Springer Berlin / Heidelberg, 2010.

[11] A. Roberts, C. Trapnell, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, Submitted.

[12] L. Shi, L. Reid, W. Jones, R. Shippy, J. Warrington, S. Baker, P. Collins, F. de Longueville, E. Kawakasi, K. Lee, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

[13] S. Srivastava and L. Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38:e170, 2010.

[14] C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, GK, M. van Baren, S. Salzberg, B. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, May 2010.