BIO-463: Report

Antoine Masanet (IC)

***Abstract*— We analyze the paper of Chae, Danko & Kraus and reproduce one of its figures using their transcription-unit detection tool called groHMM on human cells. Furthermore, we use their tool on two other very different organisms: *Drosophila melanogaster* and *Arabidopsis thaliana* cells and compare the results and characteristics of the trained model.**

I. INTRODUCTION

The following report is based on the paper *groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data* [1]. In this paper, the authors have developed a Hidden Markov Model that uses Global Run-On Sequencing (GRO-Seq) data to predict the location of transcription units in the genome. This report will analyze their model, reproduce some of their results as well as use the model on GRO-seq data different from the one presented in the paper.

A. Transcription

Transcription is the process of copying a segment of DNA into RNA. In case the produced RNA molecules encode for a protein, we call them messenger RNA. Other RNA molecules are called non-coding RNAs. A transcription unit is the DNA segment that is transcribed, it contains the coding sequence, that will be transcribed into RNA, as well as regulatory sequences which regulates the synthesis of the RNA.

B. GRO-seq

Global Run-On sequencing (GRO-seq) is a protocol used to assess real-time transcription from engaged RNA polymerase. To do so, it provides information of the location of all three RNA polymerases in cells. Similarly to well known sequencing methods such as RNA-seq and ChIP-seq, this protocol can be used to detect the location of protein-coding transcription units. It provides a quantitative level of expression that can be used by tools such as groHMM to estimate the probability of a transcription unit to be at a certain location. The main advantage of GRO-seq over the previous methods is its ability to detect nascent RNA that degrades too quickly to be detected by the other protocols. From this nascent-RNA, it is possible to detect the location of distal regulatory elements such as enhancers RNA (eRNA).

C. HMM

A hidden Markov model is a statistical model in which the running Markov process cannot be directly observed but each hidden state directly influences a measurable random variable. The goal of such model is to estimate the sequences of hidden states by looking at the observations. These models are widely used in Bioinformatics for predicting the position of certain functional region by looking at the sequence. Predictions of CPG-islands regions [2] is one of such use.

D. groHMM

The authors of the paper report that their model achieves better accuracy than well known models SICER [3] and HOMER [4] also used on GRO-seq data to detect transcription units. The model is based on a 2 hidden-state Markov chain with states *transcribed region* and *non-transcribed region*. Instead of going over each nucleotide individually, the model takes as input read counts from GRO-seq data in 50 bp windows. As an output, it classifies each of those 50 bp regions as transcribed or not (Figure 1).

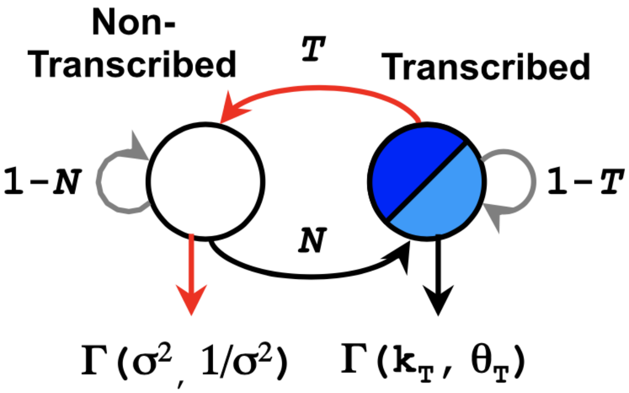

Fig. 1. The groHMM Markov Chain

As we can see, the model is described by its hidden states transition probabilities and its emission probabilities. This emission distribution can be hard to define as it may vary a lot between species or even different cells within the same species. As a result, the authors chose a gamma distribution that can model a great variety of distribution by modifying its shape and scale. This model has a total of 5 parameters:

- T: hyper-parameter used to control the length of transcription units, it models the probability of going from transcribed to non-transcribed region.
- N: probability of going from non-transcribed to transcribed region
- σ^2: hyper-parameter used to model the emission probability (number of detected reads) in the non-transcribed region. Assuming perfect data, this probability should be 0. Since the GRO-seq data is noisy, the number of read count in this non-transcribed region is modeled by a $\Gamma(\sigma^2, 1/\sigma^2)$ distribution.

– k_t and θ_t: used to model the number of read counts in the transcribed region

E. Experimental Data

Publicly available GRO-seq data sets were obtained from the NCBI GEO repository using the following accession numbers:
MCF-7: GSM678535, GSM678536, GSM678537, GSM678538, GSM678539, GSM678540, GSM678541 and GSM678542.
Drosophila melanogaster: GSM1020091, GSM1020092, GSM1020093 and GSM1020094.
Arabidopsis thaliana: GSM3682879, GSM3682880 GSM368281 and GSM3682882.

F. Code

All code used in this report is hosted on a Github repository. Everything is coded in R and Python and the conclusions can be reproduced using the previously declared experimental data.

II. Task 1

For the first task, the figure chosen to be reproduced from the original paper is shown below.

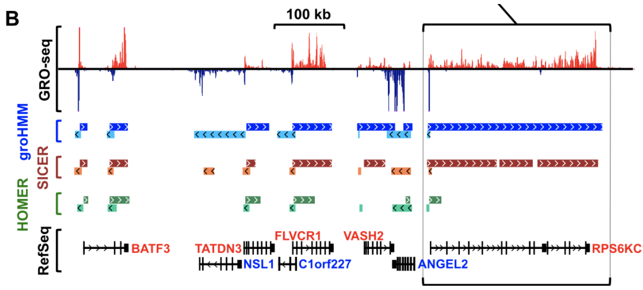

Fig. 2. Genome browser tracks of GRO-seq data from MCF-7 cells with transcription units called by groHMM and corresponding RefSeq annotations

The top row of this figure shows the raw GRO-seq data obtained from the MCF-7 breast cancer cells. The red part represent the positive strand read count and the blue part represent the negative strand read count. The second row indicates the predicted transcriptionally active genomic regions using the groHMM algorithm with arrows pointing to the right representing the positive strand and conversely for the arrow pointing the left. Finally, the bottom row shows the NCBI RefSeq annotations that indicates the current consensus for the coding region of the human genome. We only try to reproduce the output of the groHMM algorithm and not the output of the SICER and HOMER algorithms.

A. Preprocessing the data

1) LiftOver: To reproduce the figure, we first downloaded the MCF-7 cells dataset from the NCBI GEO repository as indicated in the paper. Those datasets correspond to the GRO-seq sequencing of MCF-7 cell lines at different stages of its treatment with E2 hormone. The data from those BED files where then lifted from the hg18 assembly to the hg19 assembly using the UCSC Lift-Over tool. This maps all the genomic data to the same reference genome and makes alignment with other sequences and the UCSC RefSeq coherent. It also removes the reads that could not be mapped to any chromosome in the original study indicated by the *rRNA* tag in the raw data.

2) Merging: The obtained data gave for each nucleotide, the number of reads of nascent RNA that where found at its position. We then merged the data of the 8 previously lifted BED files by adding up the read count of overlapping nucleotide ranges. Furthermore, we cleaned the remaining data by removing one line with a non-standard chromosome tag *chr11_gl000202_random* to obtain a total of 32,620,584 reads. In the supplementary material of the paper, it is specified that the read counts were doubled before being provided to the groHMM algorithm. As a result, we multiplied by 2 all read counts to obtain a total of 63,773,686 reads compared to the 63,473,424 reads of the paper, which is an 0.4% difference. This discrepancy could be explained by different parameters used in the lift-Over tool that were not specified in the paper, we used the default provided parameters for this study.

3) Splitting: At this point, the obtained data is of this form:

Chromosome	range	strand	read count
chr2	5440-5441	+	2
chr2	5446-5447	+	1
chr2	5450-5451	+	1
chr2	5467-5468	+	3

However, the functions provided by the authors to train and run the groHMM algorithm do not take into account the read count of each range but count each line of the dataframe as one read. To solve this issue, we used Python and the Panda libraries to replicate each range in the dataframe by the number of reads detected in it. The resulting dataframe is shown below:

Chromosome	range	strand
chr2	5440-5441	+
chr2	5440-5441	+
chr2	5446-5447	+
chr2	5450-5451	+
chr2	5467-5468	+
chr2	5467-5468	+
chr2	5467-5468	+

B. DetectTranscripts

Finally, we performed the training and prediction of the model using the detectTranscripts function provided with the library from the original manuscript. According to the supplementary data of the paper, we set the two initial hyper parameters $\sigma^2 = 30$ and $\log T = -350$ and trained the model. After around 30 minutes of training on a commodity computer, the algorithm outputs the ranges of nucleotides it predicted as being transcribed regions as well as the trained parameters it converged to. As a result, the algorithm marked

34,145 non-overlapping genomic regions as transcriptionally active. Here is a sample of the obtained data:

Chromosome	range	strand
chr2	5800-7099	+
chr2	9200-9549	+
chr2	19000-19499	+
chr2	21800-47599	+
chr2	52600-58499	+

C. Export data

To obtain the same figure as the paper, we had to export the data in Wiggle files to be able to visualize it on the UCSC Genome Browser. These wiggle files enabled us to see the data in fixed step of 50 bp showing how the algorithm views the data. Furthermore, we formated the files with the correct colors and visualization to be as close as possible to Figure 2. Figure 3 illustrates our results.

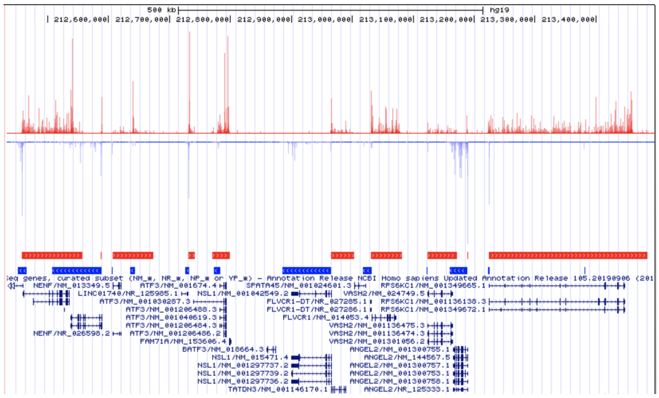

Fig. 3. Genome browser tracks of GRO-seq data from MCF-7 cells with transcription units called by groHMM (blue and red middle arrows) and corresponding RefSeq annotations (blue at the bottom)

D. Result and discussion

Overall the resulting Figure 3 is very similar to the figure we intended to reproduce (Figure 2). We can observe by looking at the second row that the predicted transcribed regions closely match the consensus annotations displayed in the third row. These specific results as well as other views of the model along the genome lead us to conclude that the model developed by the authors for detecting transcriptionally active regions in the genome is a useful and accurate tool.

III. Task 2

As we have previously seen, the model provided by the authors achieves a good accuracy for predicting transcriptionally active regions using GRO-seq data of the human genome. The question we could then ask ourselves is if this model can be generalized for other species. The paper further applies their model on *D. melanogaster* and *C.elegans* and note that the model has a tendency to merge together close transcriptionally active genomic regions. To explore this claim, we have decided to reproduce their study on publicly available GRO-seq data of *Arabidopsis thaliana* [5]. This plant has a high gene density (321.9 genes/Mb [8] compared to 15.5 genes/Mb for Human [6]) which makes it well suited for the comparison. Furthermore, we will also reproduce their study on *D. melanogaster* to obtain the HMM parameters of three trained models and be able to compare them.

A. Drosophila

Using the GRO-seq data in the paper, we trained the model for *Drosophila melanogaster*. To do so, we choose the paper's hyper parameters $\sigma^2 = 50$ and $\log T = -50$. Fig 4 illustrates a view of the obtained result:

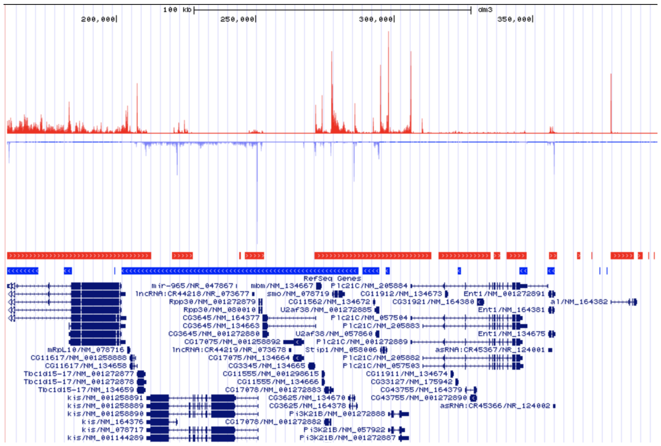

Fig. 4. Genome browser tracks of GRO-seq data from Drosophila cells with transcription units called by groHMM (blue and red middle arrows) and corresponding RefSeq annotations (blue at the bottom)

B. Arabidopsis

The GRO-seq data found to train this model comes from Arabidopsis leave cells at different stages of its response to a heat shock. To estimate the best initial hyper-parameters for the Arabidopsis HMM model, we have trained the model using various values for $\log T$ ranging from -100 to -10 by steps of 5 until we obtained a number of transcripts that was similar to the ground truth obtained from NCBI [8]. Doing so, we found $\log T = -30$ to be the best parameter. We kept $\sigma^2 = 50$ as it was the case for training the GRO-seq data of Human and Drosophila. The NCBI genome browser does not offer the possibility to plot genomic data of *A. thaliana*, we therefore used another tool called JBrowse loaded with the Arabidopsis RefSeq data from the Arabidopsis website to show the predicted transcribed regions. Fig 5 illustrates a view of the obtained result:

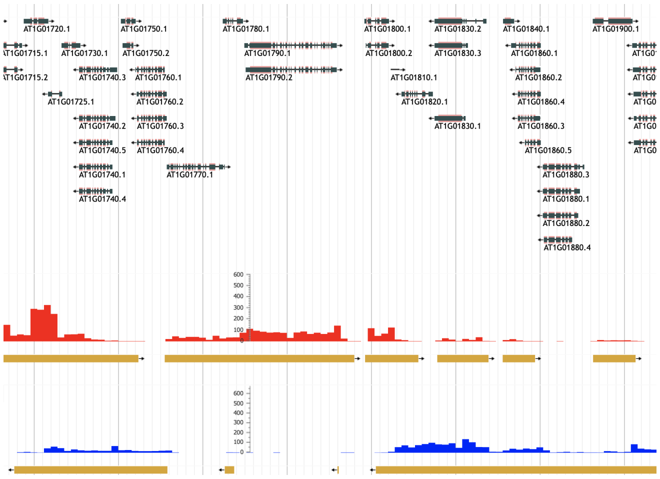

Fig. 5. JBrowser tracks of GRO-seq data from *A. thaliana* cells with transcription units called by groHMM and corresponding RefSeq annotations

This specific view was chosen to illustrate the two types of error the model makes with high gene density GRO-seq data. The first error, reported by the authors, is its tendency to merge close genomic regions into a single transcribed region that can be seen by comparing the bottom-right track and the corresponding RefSeq data. Furthermore, the model being tuned for small transcribed length, it also detects from low GRO-seq data expressions, very small transcribed regions that do not correspond to any existing gene. This can be seen at the middle of the bottom track. We will further investigate this second anomaly by plotting the distribution of predicted transcript length of Arabidopsis.

C. Model comparisons

We will now compare the trained parameters obtained from running the groHMM algorithm on the three different organisms. The table below shows us those results:

Trained parameters	Human	Drosophila	Arabidopsis
k_t	0.661	0.647	0.543
θ_t	5.723	9.518	32.426
N	$\frac{0.352}{1000}$	$\frac{3.24}{1000}$	$\frac{11.64}{1000}$

As we can see, k_t, the shape parameter of the transcribed region gamma distribution, remains very similar for the three species. On the other hand, the θ_t parameter, which represents the standard deviation of number of reads in the transcribed region, is much bigger for the model trained on Arabidopsis compared to the two other models. This increase results in a distribution with a higher probability of a transcribed region to have many reads. This difference can be explained by Arabidopsis GRO-seq data itself where more reads where detected in each transcribed region. This illustrates one of the strength of the model that can learn the varied levels of expression in the GRO-seq data without the need to normalize the data before-hand. Finally, N which determines the probability of transition from non-transcribed state to transcribed varies a lot between the species. This parameter has a direct impact on the size of the non-transcribed regions and the observed variation comes from the higher gene density of some species over others as can be seen from the table below:

Species	gene count	density (genes/Mb)
Homo sapiens [6]	44,507	15.5
Drosophila melanogaster [7]	17,868	130.4
Arabidopsis thaliana [8]	38,311	321.9

We can then compare these NCBI consensus values to the values obtained by the algorithm shown in the table below. As we can see, the model underestimated the number of transcripts for the Human and Arabidopsis genome and overestimated it for Drosophila. From this small sample of data, there does not seem to be a clear trend regarding the bias of the model as a consequence of the observed species' gene density.

Trained results	Human	Drosophila	Arabidopsis
Transcript count	34145	16123	36724
Average transcript length	32840 bp	5293 bp	2161 bp
Median transcript length	6600 bp	1450 bp	500 bp
Gene density (genes/Mb)	11.93	117.09	306.8
Transcript length Std	70617	10368	3405

Finally, we are going to compare the distribution of the predicted transcribed regions of those three organisms.

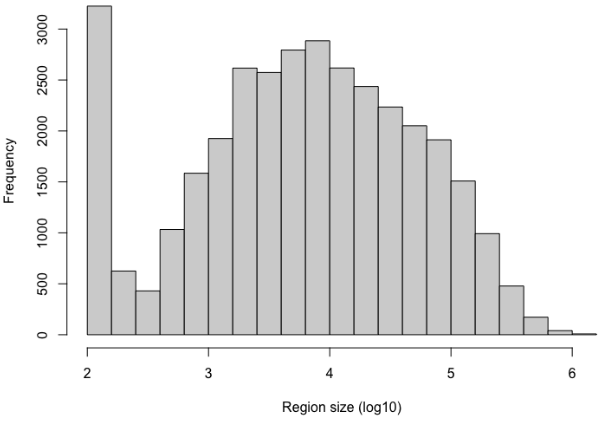

Fig. 6. Human transcript length distribution

As we have seen before, the average transcript length increase by a factor 150 between Human and Arabidopsis, as a result, we chose a logarithmic plot for the transcribed

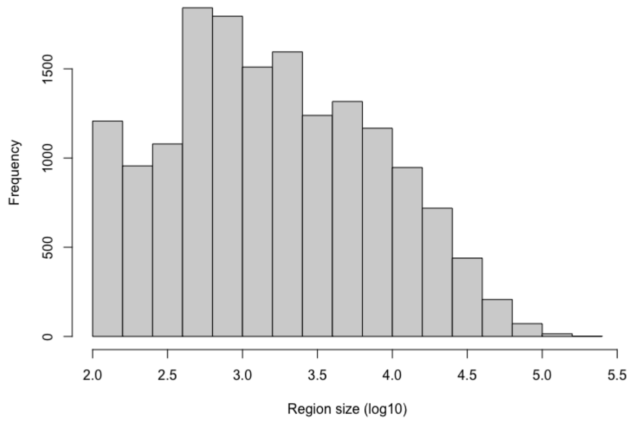

Fig. 7. Drosophila transcript length distribution

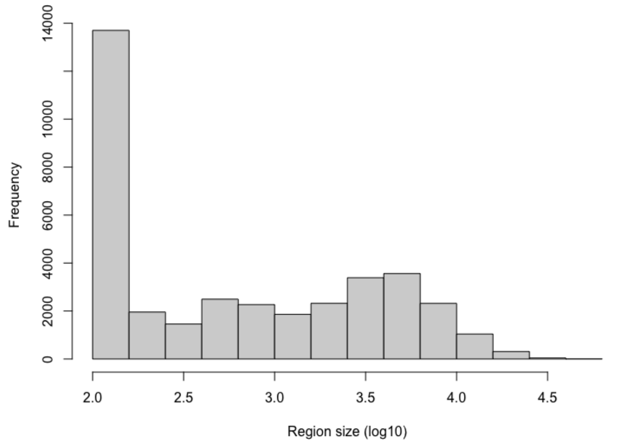

Fig. 8. Arabidopsis transcript length distribution

region size on the x-axis. Regarding the distribution of the region size, the Human and Drosophila have a similar transcriptionally active genomic region size distribution. On the other hand, the Arabidopsis groHMM model had a much higher tendency to predict small transcription regions. This high tendency of the model to predict regions of size a 100 bp (which corresponds to two 50 bp windows) in the algorithm confirms our previous observation in Fig 5.

IV. Conclusions

We have analyzed the groHMM software developed by Chae, Danko & Kraus and tested it on both their own data and for new analysis. We confirm that groHMM is a fast, simple and useful tool for predicting actively transcribed regions in the genome. The model appeared to work very well with human GRO-seq data with which we obtained similar results as the authors. We then used their model on Arabidopsis GRO-seq data and noted that the distribution obtained was very different than for lower gene density species such as the Human or Drosophila. This might be due to, as the authors point out, the high gene density of this species. Either way, the software has the advantages of being fast, free of charge and can be used to quickly gain insight in the distribution of actively transcribed regions in a species. On the downside, the high dependency of the model on the provided hyper-parameters makes it hard to study new species without already knowing the expected gene length predictions.

References

[1] Chae M, Danko CG, Kraus WL. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC Bioinformatics. 2015 Jul 16;16:222. doi: 10.1186/s12859-015-0656-3. PMID: 26173492; PMCID: PMC4502638.

[2] Hao Wu, Brian Caffo, Harris A. Jaffee, Rafael A. Irizarry, Andrew P. Feinberg, Redefining CpG islands using hidden Markov models, Biostatistics, Volume 11, Issue 3, July 2010, Pages 499–514, https://doi.org/10.1093/biostatistics/kxq005

[3] Chongzhi Zang, Dustin E. Schones, Chen Zeng, Kairong Cui, Keji Zhao, Weiqun Peng, A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, Bioinformatics, Volume 25, Issue 15, 1 August 2009, Pages 1952–1958, https://doi.org/10.1093/bioinformatics/btp340

[4] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576-89. doi: 10.1016/j.molcel.2010.05.004. PMID: 20513432; PMCID: PMC2898526.

[5] Liu M, Zhu J, Dong Z. Immediate transcriptional responses of Arabidopsis leaves to heat shock. J Integr Plant Biol 2021 Mar;63(3):468-483. PMID: 32644278

[6] National Center for Biotechnology Information (NCBI)[Internet] https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens%5Borgn%5D

[7] National Center for Biotechnology Information (NCBI)[Internet] https://www.ncbi.nlm.nih.gov/genome/?term=Drosophila_melanogaster+%5Borgn%5D

[8] National Center for Biotechnology Information (NCBI)[Internet] https://www.ncbi.nlm.nih.gov/genome/?term=Arabidopsis_thaliana+%5Borgn%5D