A Maximum Entropy Approach to Defining Cell State

Ashika-Sita Jayanthy

Abstract

The past few decades have seen great leaps in technologies to analyze cells and tissues. Omics methods in particular now allow us unprecedented access to their the molecular composition where the base-level resolution of transcripts and their numbers can be determined at a single cell level. Existing methods to analyze the resulting data make use of the count data while discarding the information present in the sequences themselves. In this paper we used a maximum entropy approach to develop a method to analyze RNA-seq data using both the sequence and count information. By mapping sequences to vectors of spins and defining an energy function on them, we were able to identify specific states in a biological process using mean energies and their associated Boltzmann-probabilities. This approach opens up new avenues in the quantitative analysis of -omics data and analysis of biological function.

Introduction

In a seminal breakthrough, Claude Shannon formalized the notion of information [1]. Since then, the idea of information in various biological systems has been explored and maximum entropy methods in particular have found broad use in several biology problems, where the probability distributions of biological processes are defined using the least structure possible for the given information [2] [3] [4]. To this end, embryo development provides a well studied framework within which to study biological information flow as embryo development is characterized by synchronized cell division and movement events tightly controlled by modules of genes expressed in precise spatial and temporal patterns [5] [6].

In this study we used the principles of maximum entropy to identify a state in a biological process using an energy function. We explored the consequences of this using simulated and experimental sequence data and found that defining states in biological processes in this manner can be surprisingly informative.

Results

Defining a biological state

A biological process such as development or cancer can be defined by its RNA composition. Since the advent of single cell sequencing, cell types are determined by the relative counts of transcripts. However, these methods treat the expressed transcripts as indistinguishable, assigning categorical labels of gene names to the counts. Since biological processes are characterized by information flow across multiple dimensions, we hypothesized that the sequence composition of the transcriptome associated with a species had underlying structure in it that we could exploit. In this case we define structure as the difference from an appropriately chosen random distribution, allowing for information flow. To test this hypothesis, we obtained sequences for the transcriptome of *Xenopus tropicalis*, a key model organism in the study of early embryo development. To probe the structure of the transcriptome, we first mapped the bases of each RNA sequence to a vector of "spins" as follows (Figure 1):

\[
S(\{\sigma_i\}) = \begin{cases} 
A & \sigma_i = 1 \\
T & \sigma_i = -1 \\
G & \sigma_i = 2 \\
C & \sigma_i = -2 
\end{cases}
\]
Next, used the simplest possible enforcement of structure: nearest-neighbour two-point correlations to define an effective energy term associated with these correlations that corresponds to an Ising-like function:

\[ E(\sigma) = \sum_i h_i(\sigma_i) + \sum_{i,j} J_{i,j}(\sigma_i \sigma_j) \]

The Boltzmann-probability of observing a sequence then depends on both the composition of the sequences, as well as the weights associated with each spin and spin-pair:

\[ p_i = \frac{\exp(-kE_i)}{\sum_j \exp(-kE_j)} \]

If we assume the observed counts of the transcripts are an estimate of their probability, a state in a biological process can then be defined by the distribution of the effective energies of its sequences.

**Inference of \( \{h, J\} \)**

Given that the transcriptome of a species is relatively constant within the timescales of an organismal process, this is a problem where the spins are unchanged, but we would like to determine the weights associated with each spin and pair of spins. We thus formulated this as an Inverse Potts problem where the goal is to learn the \( \{h, J\} \) for a set of sequences.

To infer the weights associated with a set of spins we used an adapted version of the method in [3]. Briefly, starting with an initial set of \( \{h, J\} \) assigned by us, we updated them as a function of their log fold-difference to the estimated probability distribution obtained using experimental count data (see Methods).

To test this method, we generated random sets of sequences and randomly assigned them "true" probabilities. We inferred the \( \{h, J\} \) for each sequence and calculated their probabilities. We compared the inferred and true probabilities using Pearson’s correlation coefficient (Figure 2).

**Tuning the underlying distributions of \( \{h, J\} \)**

To explore this idea further, generated random sequences with base-biases. We then drew \( \{h, J\} \) from negative binomial distributions with different parameters, and clustered cells using the resulting counts using tSNE [7] (Figure 3). We found that changing the parameters for the distributions of \( \{h, J\} \) for same set of randomly drawn base-biased sequences resulted in changing numbers of "cell clusters".

**Xenopus tropicalis embryo development**

Next we obtained single-cell RNA-sequencing data from stages 8 to 22 of the *Xenopus tropicalis* embryo development [5]. We estimated the \( \{h, J\} \) using the maximum entropy approach detailed above. We found that we were able to distinguish between stage and cell type using only the mean energies (Figure 4).

**Methods**

**Single-cell RNA-seq data processing and analysis**

Single cell RNA-seq data for the embryonic development of *Xenopus tropicalis* [5] was downloaded from the NCBI Gene Expression Omnibus (NCBI GEO) Accession ID GSE113074. Corrected, combined and annotated counts were analyzed both with and without normalization for total transcript counts per cell per stage. Counts were pooled and normalized for stage-level expression.

**Learning \( \{h, J\} \)**

The probability distribution \( P_{1\text{observed}} \) identifies the probability of observing a specific base at each position in a sequence. This was calculated by the summing the probabilities of each sequence with a base in a specific position.

Similarly, the probability of each pair correlation, \( P_{2\text{observed}} \) was calculated by summing the probabilities of each sequence with pair correlation.

\( \{h, J\} \) for each set of sequences were initialized with 1s. We calculated the effective energy of each sequence given a the estimate of \( \{h, J\} \). The probability of each sequence given its effective energy was then calculated using the Boltzmann probabilities:

\[ p_i = \frac{\exp(-kE_i)}{\sum_j \exp(-kE_j)} \]
where \( k \) was chosen to be \( 1 \times 10^{-3} \). We then calculated the probabilities of each base in each position and each pair correlation for each possible pair of positions.

We call these distributions \( P^{1}_{\text{model}} \) and \( P^{2}_{\text{model}} \). We then updated the \( \{ \mathbf{h}, \mathbf{J} \} \), using the log fold-difference between the observed and inferred probabilities:

\[
\begin{align*}
\epsilon_1 & \propto \log \left( \frac{P^{1}_{\text{observed}}}{P^{1}_{\text{model}}} \right) \\
\epsilon_2 & \propto \log \left( \frac{P^{2}_{\text{observed}}}{P^{2}_{\text{model}}} \right)
\end{align*}
\]

**Generation of base-biased sequences**

We generated base-biased sequences according to the probabilities 0.167, 0.25, 0.25, 0.333 for A, T, G, and C respectively.

**Discussion**

The analysis of various biological functions using single cell RNA-seq has taken flight in recent years [8] [9] [10], where the composition of RNA in cells is used to define cell type and then track their changes over time. These methods rely on using count matrices for genes, and cells are clustered into various categories based on the relative counts of every gene measured using dimensionality reduction methods such as tSNE and UMAP [7] [11]. While useful in a qualitative sense, these clusters offer little by the way of testable, quantitative predictions. Our method provides a novel quantitative framework to defining cell states using sequencing data, which potentially allows us to move beyond qualitative classifications of cell types. Future work will more fully explore the relationships between the underlying distributions of the spin coefficients and spins to generate different biological processes.

**Conclusion**

In this study we mapped sequences to spin vectors and defined an energy function on them using nearest-neighbour two-point correlations. We then used a Boltzmann-like distribution to estimate the probabilities associated with a specific set of sequences defining a biological state and found that this simple enforcement of structure on RNA sequences was enough to discriminate between different cell types and states. This is a relatively surprising result considering the dimensionality and complexity of these data.

**References**

[1] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4):623–656, 1948.

[2] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[3] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(12):5405–5410, 2010.

[4] Daniel L. Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

[5] James A Briggs, Caleb Weinreb, D.E. Wagner, Sean Megason, Leonid Peshkin, M.W. Kirschner, and A.M. Klein. The dynamics of gene expression in vertebrate embryogenesis. *Science*, 360(6392), 2018.

[6] Janet Heasman. Patterning the early Xenopus embryo. *Development*, 133(7):1205–1217, 2006.

[7] Visualizing Data using t-SNE. *Journal of Machine Learning Research*, (9):2579–2605, 2008.

[8] Mario L. Suvà and Itay Tirosh. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell*, 75:7–12, 2019.
[9] Laura González-Silva, Laura Quevedo, and Ignacio Varela. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. *Trends in Cancer*, 6(1):13–19, 2020.

[10] Kyle J. Travaglini, Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V. Sit, Stephen Chang, Stephanie D. Conley, Yasuo Mori, Jun Seita, Gerald J. Berry, Joseph B. Shrager, Ross J. Metzger, Christin S. Kuo, Norma Neff, Irving L. Weissman, Stephen R. Quake, and Mark A. Krasnow. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, 587(7835):619–625, 2020.

[11] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2020.
Figure 1: Defining a cell state using RNA composition
Figure 2: Histogram of Pearson's correlation coefficient between known and inferred probabilities
Figure 3: t-SNE plots of simulated sequencing data

Figure 4: Inferred effective energy of cell types in *X.tropicalis* development