MINIMAL ERROR CORRECTING GENETIC CODE FOR NEURAL IDENTITY IN *C. elegans*

Molly B. Reilly, Cyril Cros, Erdem Varol, Eviatar Yemini, and Oliver Hobert

May 5, 2020

1 Introduction

Given a codebook of genetic expression in a population of cells, we explore the problem of whether it is possible to reduce this codebook to its minimal form such that no two cells possess an identical expression of genes. This is in contrast with the problem of optimal coding in information-theoretic contexts where techniques such as Huffman coding [1] aim to encode a set of symbols with variable observation probabilities using a set of bits such that the expected codeword length is minimized. In the genetic context, we already observe a given set of redundant codes for each cell that is the result of differentiation processes and we aim to reduce this codebook where there are no redundancies and each cell is represented by a unique barcode. The problem of codebook reduction is cast as a multidimensional knapsack problem [2] with binary weight constraints. The global optimum solution is then found through a branch-and-bound scheme [3] that yields the minimum subset of bits that can be conserved from the genetic codebook that ensures uniqueness of cell barcodes.

2 Method

First, we introduce notation. Let $\boldsymbol{T} \in \{0,1\}^{n,k}$ denote the binary matrix of gene expression of n cells with k genes. we assume that the initial set of genes is an overcomplete barcode such that no two cells have identical gene expression. Symbolically, if $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ denote the ith and jth rows of the gene expression matrix corresponding to the ith and jth cell, then we have that $\mathbf{1}^T|(\boldsymbol{t}_i - \boldsymbol{t}_j)| \geq 1$. Given this set of gene expressions, we aim to find a minimal set of genes, encoded by the vector $\boldsymbol{w} \in \{0,1\}^k$ such that under this subset of genes, no two cells have identical expression patterns. Symbolically, this can be written as $\boldsymbol{w}^T|(\boldsymbol{t}_i - \boldsymbol{t}_j)| \geq 1$. Combining these criteria and constraints yields the following objective function:

$$\begin{aligned} &\text{minimize } \ \boldsymbol{w}^T\mathbf{1} \\ &\text{subject to} \\ &\boldsymbol{w} \in \{0,1\}^k \\ &\boldsymbol{w}^T|(\boldsymbol{t}_i - \boldsymbol{t}_j)| \geq 1, \ \ \forall i \neq j \end{aligned} \tag{1}$$

This problem is an NP-complete problem, being a mixed-integer linear program. However, the global optimum can be obtained through branch-and-bound[4, 5] where we can branch on the bits that can be removed and solve the lower bound of each node by solving the relaxed linear program:

$$\begin{aligned} &\text{minimize } \ \boldsymbol{w}^T\mathbf{1} \\ &\text{subject to} \\ &0 \leq \boldsymbol{w} \leq 1 \\ &\boldsymbol{w}^T|(\boldsymbol{t}_i - \boldsymbol{t}_j)| \geq 1, \ \ \forall i \neq j, \forall i \in NodeSet \end{aligned} \tag{2}$$

3 Results

We ran our algorithm on the expression table of 118 neuron classes of *C. elegans* with 63 conserved homeodomain transcription factors expressions. The expression patterns of the neuron classes are illustrated in the heatmap in figure 1.

The branch-and-bound algorithm converged to a solution set of 24 transcription factors illustrated in figure 2.

3.1 Error correcting codes

Once the minimal codebook has been established, it is possible to include additional error correcting redundancies to account for biological variability in transcription. This can be done by computing the correlation of the minimal codebook genes with the remaining gene set and choosing the most correlated genes. This is illustrated in figure 3 where the expression correlation of additional transcription factors to the minimal codebook transcription factors are shown.

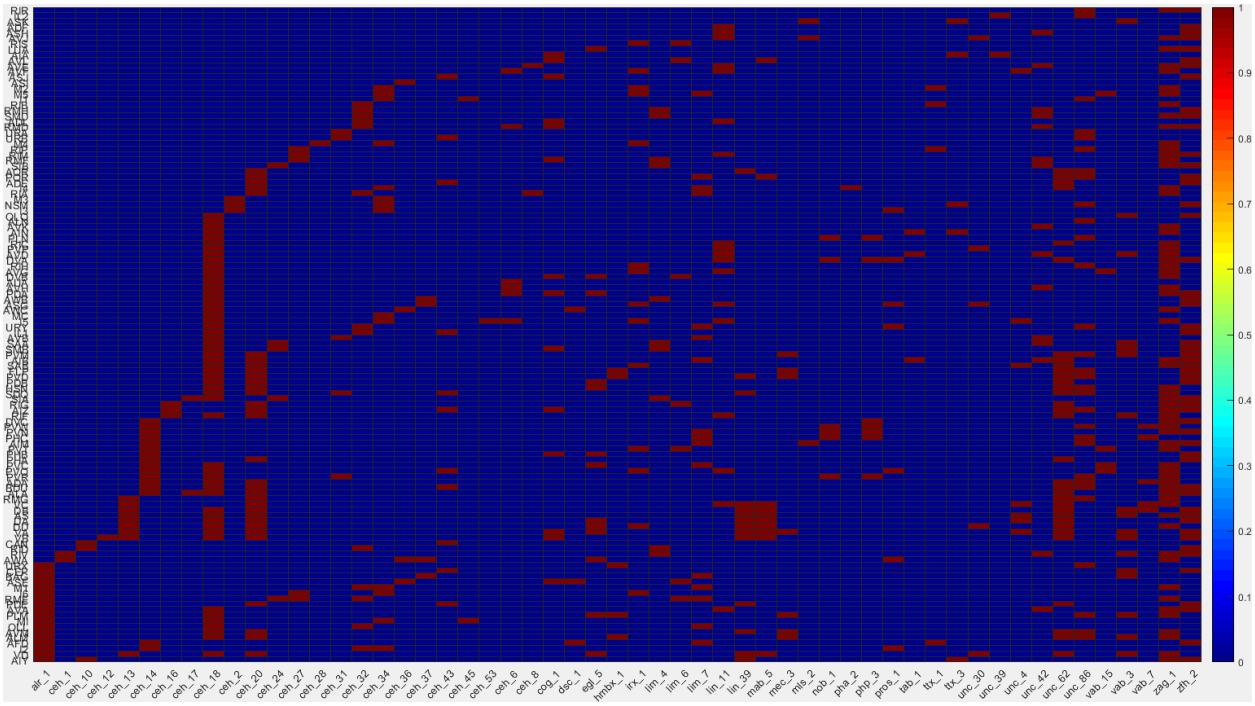

Figure 1: The homeodomain transcription factor codebook of *C. elegans* neuron classes. Rows: neuron classes, columns: transcription factor expression

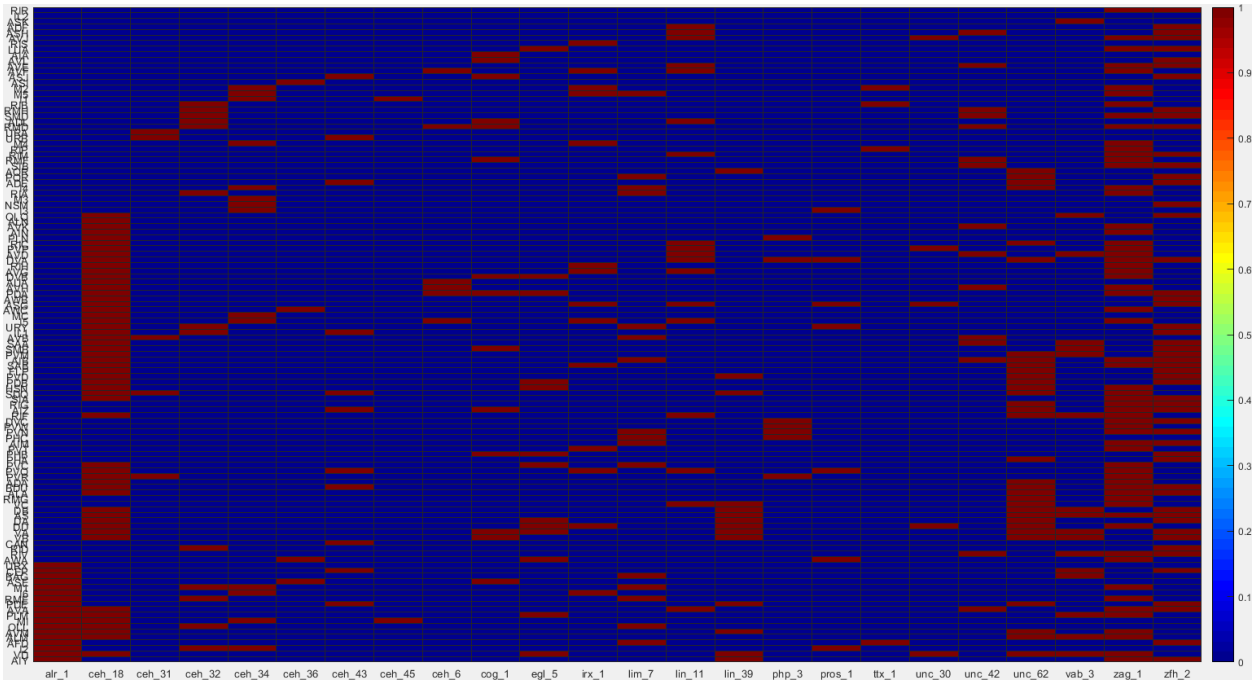

Figure 2: The minimum codebook of transcription factor barcodes that unique identifies each neuron class in *C. elegans*, Rows: neuron classes, columns: transcription factor expression

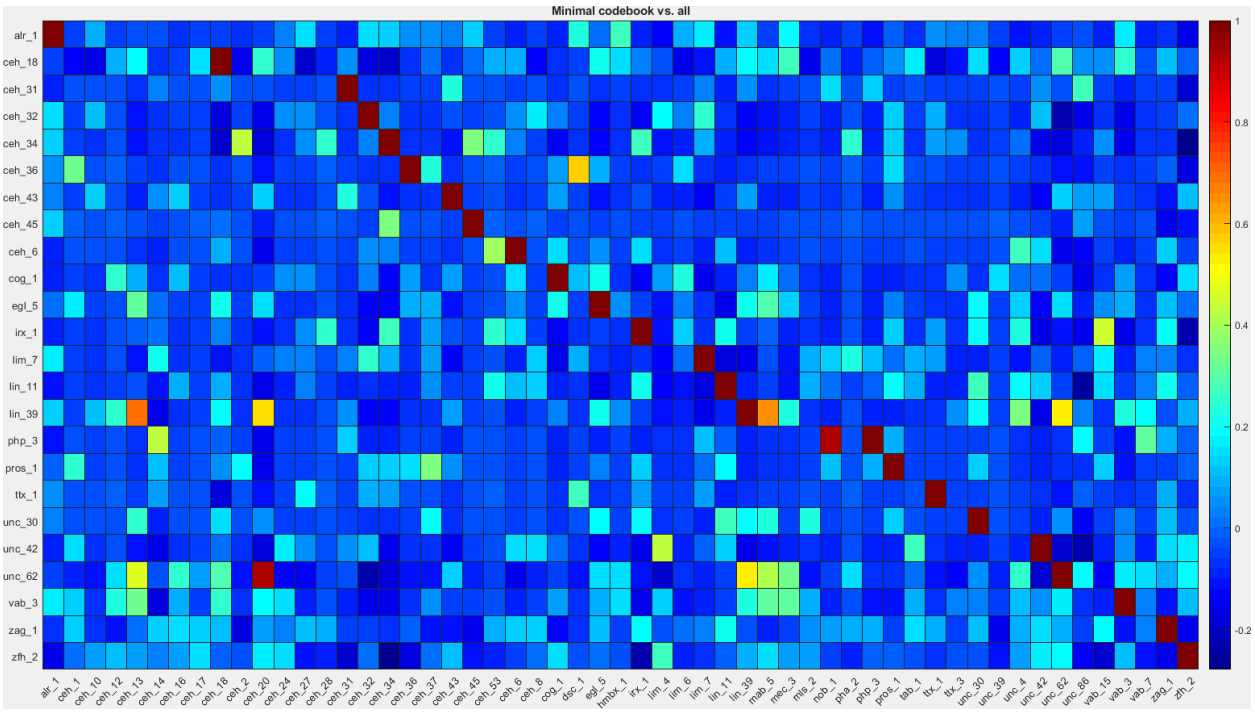

Figure 3: The candidates (columns) for error correcting redundant bits can, be selected based on correlation to the minimal codebook bits (rows).

References

[1] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[2] Hans Kellerer, Ulrich Pferschy, and David Pisinger. Multidimensional knapsack problems. In *Knapsack problems*, pages 235–283. Springer, 2004.

[3] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[4] Nicolas Nadisic, Arnaud Vandaele, Nicolas Gillis, and Jeremy E Cohen. Exact sparse nonnegative least squares. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5395–5399. IEEE, 2020.

[5] Ramzi Mhenni, Sébastien Bourguignon, and Jordan Ninin. Global optimization for sparse solution of least squares problems. 2019.