iDeLUCS: A deep learning interactive tool for alignment-free clustering of DNA sequences

Pablo Millán Arias ¹*, Kathleen A. Hill², Lila Kari¹

¹ School of Computer Science, University of Waterloo, Waterloo, ON, Canada
² Department of Biology, University of Western Ontario, London, ON, Canada

* pmillana@uwaterloo.ca

Abstract

Summary: We present an interactive Deep Learning-based software tool for Unsupervised Clustering of DNA Sequences (iDeLUCS), that detects genomic signatures and uses them to cluster DNA sequences, without the need for sequence alignment or taxonomic identifiers. iDeLUCS is scalable and user-friendly: Its graphical user interface, with support for hardware acceleration, allows the practitioner to fine-tune the different hyper-parameters involved in the training process without requiring extensive knowledge of deep learning. The performance of iDeLUCS was evaluated on a diverse set of datasets: several real genomic datasets from organisms in kingdoms Animalia, Protista, Fungi, Bacteria, and Archaea, three datasets of viral genomes, a dataset of simulated metagenomic reads from microbial genomes, and multiple datasets of synthetic DNA sequences. The performance of iDeLUCS was compared to that of two classical clustering algorithms (k-means++ and GMM) and two clustering algorithms specialized in DNA sequences (MeShClust v3.0 and DeLUCS), using both intrinsic cluster evaluation metrics and external evaluation metrics. In terms of unsupervised clustering accuracy, iDeLUCS outperforms the two classical algorithms by an average of ∼20%, and the two specialized algorithms by an average of ∼12%, on the datasets of real DNA sequences analyzed. Overall, our results indicate that iDeLUCS is a robust clustering method suitable for the clustering of large and diverse datasets of unlabelled DNA sequences.

Availability and implementation: iDeLUCS is available at our github repository under the terms of the MIT licence.

Contact: pmillana@uwaterloo.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

A Introduction

Clustering algorithms for DNA sequences play a fundamental role in bioinformatics, as they can be used to study the structural composition of DNA sequence datasets, to discover novel operational taxonomic units (OTU), and to complement phylogenetic analysis. The development of high throughput sequencing technologies has raised several challenges to many clustering methodologies, as
most of them cannot keep up with the exponential increase in the number of sequences available for analysis. One of the reasons is that many clustering methods rely on the computationally expensive process of sequence alignment. To address these limitations, several alignment-assisted and alignment-free methodologies were proposed, see, e.g., [4, 2]. The majority of these methods also face scalability issues, as they are reliant on classic clustering algorithms that perform well in the low data regime but have poor performance when large amounts of data are available. While in the aforementioned approaches the exponential increase of data is a hindrance to good performance, other approaches, e.g., deep-learning-based methods, benefit from the availability of large amounts of data. In particular, multiple deep learning algorithms have been developed for classification and inference using both alignment-based and alignment-free methodologies [10, 9, 8]. It has also been shown recently by [7] that deep learning provides a significant improvement over classical unsupervised learning algorithms in discovering genomic-signature-based clusters, at different taxonomic levels. These promising initial results motivated the development of iDeLUCS, which takes advantage of the capabilities of deep learning, and is capable of clustering datasets comprising more than 400 Mbp. In addition, iDeLUCS exhibits several novel features which enhance the interpretability of its results: confidence scores of the final cluster assignments, a graphical user interface (GUI), dynamic visualization of the underlying training process, and incorporated evaluation metrics.

B Software Description

iDeLUCS is a standalone software tool that exploits the capabilities of deep learning to cluster genomic sequences. It is agnostic to the data source, making it suitable for genomic sequences taken from any organism in any kingdom of life. iDeLUCS assigns a cluster identifier to every DNA sequence present in a dataset, while incorporating several built-in visualization tools that provide insights into the underlying training process and the composition of the datasets (Figure 1). iDeLUCS offers an evaluation mode to compare the ground-truth label assignments (or hypothesized label assignments) of the dataset sequences with their discovered cluster labels. This is accompanied by a visual qualitative assessment of the clustering, through the use of the uniform manifold approximation (UMAP, see [6]) of the learned lower dimensional embedding. Finally, iDeLUCS outputs confidence scores for all of its cluster-label predictions, for enhanced interpretability. The software was developed using Python 3.9 and can be deployed with or without a graphics processing unit (GPU) (see Appendix A for implementation details).

C Materials and Methods

iDeLUCS builds upon the pipeline proposed in [7], consisting of: (i) calculating the k-mer frequencies for each DNA sequence, (ii) computing the data augmentations (mimic sequences), (iii) training multiple deep neural networks to learn
Figure 1: Training tab of iDeLUCS. The left panel displays a summary of the main training parameters, as well as some statistics about the dataset under study. The center panel contains a qualitative assessment of the learning progress. In this particular case, the figure illustrates the clustering of mitochondrial genomes of insects into 7 different clusters, each corresponding to one corner of the heptagon. Each point represents a genome, and its position indicates the probability that it is assigned to a different cluster/corner. The right panel contains a dynamic plot with the learning curves of the different models and serves as an indicator of whether or not the contrastive loss function is being minimized during training.
Table 1: Description of the new mitochondrial DNA datasets (b). Note that there is a balanced version of each new dataset (Fungi, Protists, Insects). For the balanced version, the number of sequences per cluster was selected according to the number of sequences available in the smallest cluster.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total no. sequences</th>
<th>Min. seq. length (bp)</th>
<th>Avg. seq. length (bp)</th>
<th>Max. seq. length (bp)</th>
<th>Total no. clusters</th>
<th>Cluster min. size</th>
<th>Cluster avg. size</th>
<th>Cluster max. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insects</td>
<td>9,027</td>
<td>14,602</td>
<td>15,841</td>
<td>26,613</td>
<td>7</td>
<td>652</td>
<td>1,290</td>
<td>1,976</td>
</tr>
<tr>
<td>Fungi</td>
<td>1,759</td>
<td>20,063</td>
<td>62,644</td>
<td>99,976</td>
<td>3</td>
<td>335</td>
<td>586</td>
<td>889</td>
</tr>
<tr>
<td>Protists</td>
<td>2,581</td>
<td>5,493</td>
<td>17,141</td>
<td>69,503</td>
<td>3</td>
<td>315</td>
<td>860</td>
<td>1,642</td>
</tr>
<tr>
<td>Insects–balanced</td>
<td>4,550</td>
<td>14,602</td>
<td>15,897</td>
<td>25,011</td>
<td>7</td>
<td>650</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td>Fungi–balanced</td>
<td>1,005</td>
<td>21,684</td>
<td>60,657</td>
<td>99,976</td>
<td>3</td>
<td>335</td>
<td>335</td>
<td>335</td>
</tr>
<tr>
<td>Protists–balanced</td>
<td>945</td>
<td>5,498</td>
<td>24,697</td>
<td>69,503</td>
<td>3</td>
<td>315</td>
<td>315</td>
<td>315</td>
</tr>
</tbody>
</table>

the cluster assignments, and (iv) computing the majority voting cluster assignment for each sequence. In addition to multiple algorithmic optimizations to the pipeline, iDeLUCS significantly extends it in four main aspects. First, it uses the contrastive learning framework introduced by [1], and incorporates an additional contrastive term in the loss function, which enforces the consistency of the hidden representations learned by the artificial neural networks. These hidden representations are learned simultaneously with the cluster assignments via backpropagation. Second, it replaces the majority voting scheme by a more robust clustering ensemble based on information theory, which reduces the variance and boosts the accuracy. Third, it uses the information provided by the ensemble and the consistency of the hidden representations to provide an intrinsic quantitative assessment of the clustering assignment (silhouette coefficient, Davies-Bouldin Index), as well as to output the confidence score for the cluster assignment of each sequence in the dataset. Finally, the new contrastive learning framework can be combined with non-parametric clustering algorithms, such as HDBSCAN [5], to automatically determine the number of clusters. This iDeLUCS option is recommended only for fine-grained clusterings, due to the fact that HDBSCAN is a density-based method (see Appendix B for details).

To assess the performance and applicability of iDeLUCS, we first analyzed 14 real datasets with known ground-truth annotations (described in (a) and (b)): Nine datasets of mitochondrial DNA from various Kingdoms of life (Animalia, Protista, Fungi) totalling 18,810 sequences, two Bacteria datasets totalling 4,800 sequences, and three viral datasets totalling 4,144 sequences. In addition, we analyzed one dataset of simulated reads from microbial genomes (Bacteria and Archaea) comprising 432,333 reads (described in (c)), and 12 synthetic datasets totalling 246,625 artificial DNA sequences (described in (d)). Each dataset was selected for its unique characteristics, as described herein:

(a) Eight datasets from Kingdom Animalia, Kingdom Bacteria, and three datasets of viral sequences, obtained from [7]: six mitochondrial DNA datasets of vertebrates at taxonomic levels from Subphylum to Family; two bacterial datasets to be clustered into families; and three viral datasets (Dengue, Influenza-A, Hepatitis B) to be clustered into virus subtypes. The maximum number of clusters per dataset is 12, and the maximum
cluster size is 500 sequences, with the average sequence length of 16,700 bp for mtDNA, 433,882 bp for bacterial, and 5,058 bp for viral sequences.

(b) Three new mitochondrial DNA datasets (Table 1) created to enhance the representation across Kingdoms of life: A dataset of 2,581 mitochondrial genomes from Kingdom Protista (average sequence length 17,141 bp) clustered into three phyla/subphyla; a dataset of 9,027 mitochondrial genomes from class Insecta (average sequence length 15,841 bp) clustered into seven orders; and a dataset of 1,759 mitochondrial genomes from Kingdom Fungi (average sequence length 62,644 bp), clustered into three phyla/subphyla.

(c) One dataset of simulated metagenomic reads from eight microbial genomes, obtained from [11]. This dataset comprises more than 430,000 reads to be clustered into eight species (seven Bacteria and one Archaea). The reads were simulated using the PacBio sequencing simulation parameters, with maximum cluster size of 119,330 sequences, and average sequence length of 8,511 bp.

(d) 12 synthetic datasets from [3]. These are artificial datasets, each consisting of 100 random template sequences, representing the true clusters, and a random number of mutated copies that were generated from each template according to a pre-defined identity threshold. Each dataset contains at most 25,000 sequences, with a minimum dataset size of 18,210. The maximum number of clusters for each dataset is 12, the maximum cluster size is 400 sequences, and the average sequence length is 20,552 bp.

A detailed description of the datasets can be found in the Tables 1, 2, and 3 in Appendix C.

To assess the performance of iDeLUCS on all the datasets analyzed in this study, we utilize both intrinsic (Davies-Bouldin Index, Silhouette Coefficient) and external (Homogeneity, Completeness, Unsupervised Clustering Accuracy (ACC)) clustering evaluation metrics. Of these, ACC arguably is the best indicator of performance, as it reflects the correspondence between cluster assignments and the ground-truth.

D Applications and Results

The performance of iDeLUCS was compared against two classic clustering algorithms, \emph{k}-means++ and Gaussian Mixture Models (GMM), as well as two recent clustering methods specific to DNA sequences, DeLUCS \cite{7} and MeShClust v3.0 \cite{3}. The performance results for all four algorithms, in terms of intrinsic and external evaluation metrics as well as running time and memory usage, on the new mitochondrial datasets (b), are summarized in Table 2. The performance results for all algorithms on the other, previously published, datasets can be found in Tables S4, S5, S6 and S7 in Appendix C.

Overall, iDeLUCS has a robust performance across these very different types of datasets: small (113 sequences) or large (432,000 reads); real, simulated, or
synthetic; at different taxonomic levels ranging from phyla to subtypes; with balanced clusters or with unbalanced clusters; with cluster number varying from 3 to 100 clusters; comprising long sequences (500,000 bp) or short sequences (650 bp); consisting of homologous sequences or of non-homologous sequences. On these datasets, the iDeLUCS clustering accuracy (ACC) ranges from 78% to 100%, with an average accuracy of 90%.

In particular, iDeLUCS outperforms the other four clustering algorithms on the real datasets in (a) and (b), most of which consist of non-homologous sequences. For example, for the mitochondrial genome datasets the average accuracy (ACC) of iDeLUCS is 92.4%, while classic clustering algorithms obtain an average accuracy of 70.2%, and the second best performant algorithm has an average accuracy of only 75.42%.

Table 2: Comparison of the performance of iDeLUCS against k-means++, GMM, DeLUCS and MeShClust v3.0 clustering algorithms on the new mtDNA datasets (b), using intrinsic cluster evaluation metrics (Davies-Bouldin index, Silhouette coefficient) and external evaluation metrics (Homogeneity, Completeness, unsupervised accuracy ACC), as well as time and memory. Boldface indicates the best result, where (↑) indicates that higher is better, and (↓) indicates that lower is better. The top half of the table displays the results for the original datasets, and the bottom half those for the corresponding datasets with balanced clusters.
As seen in Table 2, for the new mitochondrial datasets (b), iDeLUCS obtains unsupervised classification accuracies ranging from 78% to 89.7%. Specifically, iDeLUCS outperforms all the other clustering algorithms for the balanced and unbalanced versions of the Insects and Fungi datasets, and has a comparable performance with the other classifiers for both the balanced and unbalanced versions of the Protist dataset. Note that the improved clustering ensemble and the new contrastive loss function of iDeLUCS significantly enhance its capability to cluster unbalanced datasets, compared to DeLUCS. These improvements become apparent in the clustering of the dataset of simulated long metage-nomic reads (c), where the accuracy of iDeLUCS is ~16% higher than that of DeLUCS, and ~7% higher than that of k-means++. For the synthetic datasets in (d), iDeLUCS obtains an average accuracy of 98.5% when the number of clusters is given as a parameter, and of 97.3% when the option of using HDBSCAN is selected to automatically determine the number of clusters. This is slightly lower than, but comparable to, the performance of MeShClust v3.0, which achieves an average accuracy of 99.3% for these synthetic datasets. Note that not all synthetic datasets analyzed in were included in this comparison, since iDeLUCS was not optimized for some types of datasets. In particular, due to existing restrictions on the size of the output layer in deep learning models, the synthetic datasets where the expected number of clusters was large (5,000) were excluded. In addition, since the real genomic datasets analyzed in this study do not include short sequences (< 500 bp), the synthetic datasets with average sequence length < 500 bp were also excluded from this comparison. Future work is needed to systematically test and optimize iDeLUCS for datasets with short reads, or datasets where more than 200 clusters are expected.

All computational tests were performed on one of the nodes of the Beluga cluster of the Digital Research Alliance of Canada (2 x Intel Gold 6148 Skylake @ 2.4 GHz CPU, 32 GB RAM) with NVIDIA V100SXM2 (16GBh memory).

In summary, this study shows that iDeLUCS outperforms other algorithms in clustering sizeable datasets of unlabelled DNA sequences, especially when homology may or may not be present, and when the user has some prior knowledge of the expected number of clusters. Overall, our analysis shows that iDeLUCS is an accurate and scalable clustering method, performant on datasets of long, homology-free DNA sequences, not tractable via alignment-based methods due to either lack of alignment or excessive time complexity.

Funding

This work has been supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grants [R2824A01 to L.K and R3511A12 to K.A.H], and Compute Canada Research Platforms & Portals Grant [616 to K.A.H].
Acknowledgements

The authors thank Shane Ding for his assistance with the testing and release of the software, Daniel Olteanu, Joseph Butler, and Connor Holmes for testing the software, and Zihao Wang for providing access to additional computational resources.
References


