

Ranking labs-of-origin for genetically engineered DNA using Metric Learning

Anonymous Author(s)*

ABSTRACT

With the constant advancements of genetic engineering, a common concern is to be able to identify the lab-of-origin of genetically engineered DNA sequences. For that reason, AltLabs has hosted the Genetic Engineering Attribution Challenge to gather many teams to propose new tools to solve this problem. Here we show our proposed method that aims to rank the most likely labs-of-origin and generate embeddings for DNA sequences and labs. These embeddings can also be used to perform various other tasks, like clustering both DNA sequences and labs and also using them as features for Machine Learning models applied to solve other problems. We will show that our method outperforms the classic training method for this task while generating other useful information

CCS CONCEPTS

• **Information systems** → **Similarity measures; Learning to rank.**

KEYWORDS

lab-of-origin, genetic engineering attribution, metric learning, triplet network, deep learning, dna, rna, plasmid

1 INTRODUCTION

AltLabs hosted the Genetic Engineering Attribution Challenge [1], providing the participants with a dataset of 63,017 DNA sequences (with their phenotype characteristics) designed by 1,314 different labs. The goal was to design machine learning models to, given a DNA sequence, identify the most likely lab-of-origin. Since it is tough to predict the correct lab because they might use similar design techniques, they decided to evaluate the solutions using top ten accuracy [2]. It means that the model needs to place the correct lab-of-origin within the ten most likely labs, according to the scores. It turns this problem into a ranking problem, instead of a simple classification, which might ask for different techniques.

During the competition, we decided to create two branches. The first is a traditional approach using classification models (with a softmax to output probabilities for each lab). The second is the use of ranking models through Metric Learning [10] (more specifically, Triplet Networks [8]). This one has the goal of learning how to extract embeddings¹ (also known as feature vectors) from DNA sequences and, at the same time, learn the embeddings of the labs. A similarity measure (cosine similarity) is then applied to generate the score between a pair of a DNA sequence and a lab. This score, instead of trying to mimic a probability, describes how similar they are.

¹Embedding is a vectorial representation of some entity composed of latent features. Humans cannot easily understand these features, but they are very useful for machine learning models. When we train a model to extract embeddings, the goal is to place similar entities closer together in the latent space.

Despite the difference in how we train those two kinds of models, we designed them very similarly. Both of them shared the same preprocessing steps, and the first layers were almost the same. They diverged a little bit because some techniques worked better for one than the other. Nevertheless, the overall structure of these models was mostly the same. Even though they were very similar, our triplet networks consistently outperformed our classification models, indicating they are more suitable for ranking tasks.

Furthermore, it is not only about raw accuracy. Suppose a new sequence from an unknown lab goes through a regular classification model. In that case, it will probably attribute it to one of the known labs while giving no clue about its uncertainty. It is a known problem and has encouraged research in an area called Out-Of-Distribution Detection [4][12]. On the other hand, our triplet network is more robust to this problem. The embedding extracted from such a sequence might differ significantly from all the labs' embeddings, generating low similarity scores. This way, we could set a threshold and decide that it comes from an unknown lab if the score is below it.

Moreover, there is also an embedding learned for an unseen lab. We train the model to push it away from all the known labs. So, suppose the new sequence's embedding is more similar to this one. In that case, we can also assume an out-of-distribution sequence. In summary, our proposed solution has a high accuracy while also dealing with new labs and providing embeddings for clustering and other tasks.

2 PROPOSED METHOD

2.1 Implementation Details

Our proposed method consists of different preprocessing techniques proved to be efficient for DNA sequences in our experiments, along with a Neural Network. As a preprocessing step, we used Byte Pair Encoding (BPE) algorithm [5] to compress the DNA sequences. It works by looking for common patterns and unifying them into tokens, increasing the vocabulary while reducing the sequences' size. It turned our vocabulary of 4 DNA bases into 1001 different tokens.

To deal with the fact that DNA sequences are circular, we used a circular shift data augmentation. This augmentation was used during training and inference, doing what is known as Test Time Augmentation. It means we feed the model with multiple versions of the same sequence (randomly circular shifted). We then take the average of the outputs. Further, to decrease the model's complexity, we also limited the sequence to 1000 tokens. This operation runs after the circular shift, so the model receives different 1000 tokens sequences per execution and combines the results.

Both types of models are composed of a Convolutional Neural Network with multiple kernels of different sizes, as proposed by Kim. We use it to extract features from the sequence and concatenate them with the binary features provided in the dataset. The

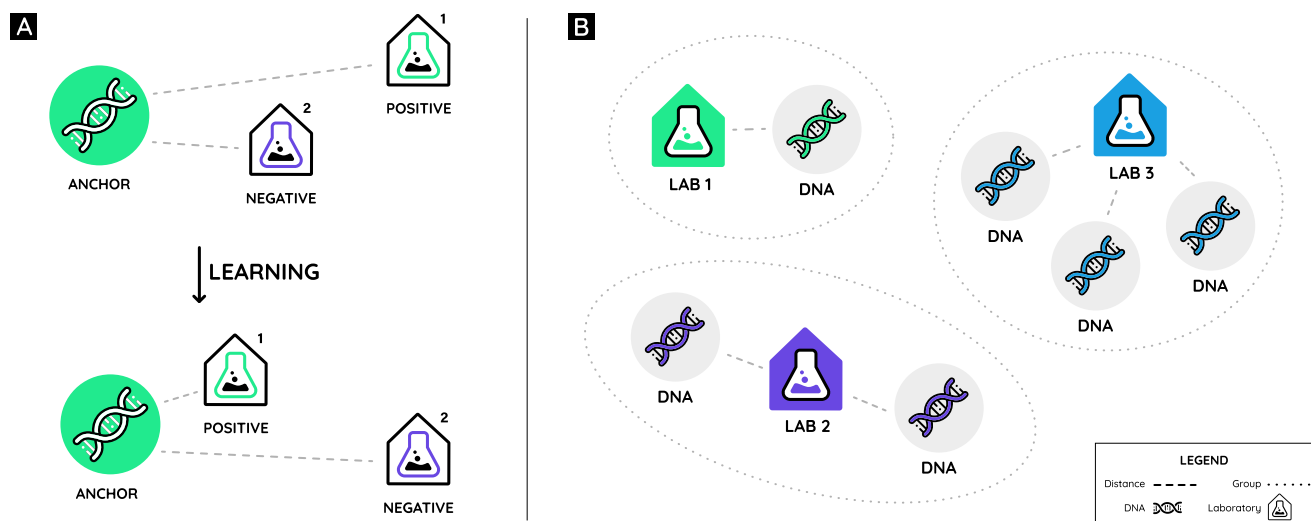


Figure 1: (A) In the beginning, the negative (another lab) might be closer to the anchor (DNA) than the positive (lab-of-origin). During training, we pull the anchor and lab-of-origin towards each other while pushing the negative away. This way, the positive will be closer to the anchor, and the negative will be farther. (B) In the end, labs and their DNA sequences will be closer to each other, forming groups. We can also expect both labs and DNA sequences to be closer to other similar ones.

difference between the classification and triplet network models are after this part.

The classification model takes these features and pass them through a dense layer and then the output layer. This output layer is a dense layer with softmax as the activation function, resulting in probabilities for each lab. On the other hand, the triplet network passes these features through a dense layer that will generate our sequence embedding. In parallel, we have an embedding layer that will learn the lab embeddings. For last, we use a cosine similarity measure between the embeddings to give us the final output.

While we train the classification model using regular Supervised Learning, we train the Triplet Network differently. We create triplets of anchors, positives, and negatives and use them for training the model. In a Face Recognition task, the anchor would be one picture of one person, the positive would be another picture of the same person, and the negative would be any picture of another person. The model learns to pull the anchor and positive together (making pictures of the same person produce similar embeddings) while pushing the negative away. In our case, the DNA sequence is the anchor, while the positive is the source lab, and the negative is another lab. Our goal is to generate embeddings in which the DNA sequences are close to their source lab and far from other labs and their sequences. Figure 1 summarizes this process.

To generate these triplets, we use the dataset to provide us with the anchor (sequence) and positive (source lab). We then use a technique known as Hard Negative Mining [6] to select the negative (another lab). It means that instead of selecting a random lab as a negative example, we select the most challenging example. It would be the lab that is currently closer to our sequence in the latent space.

2.2 Triplet Network Advantages

After training, we can use the Triplet Network to extract embeddings from DNA sequences. It also contains a table of embeddings for each lab present in the dataset and one embedding for an "unseen" lab. These embeddings can be very useful beyond the expected usage (ranking possible labs-of-origin given a DNA sequence). We can also use them to compare and cluster DNA sequences and labs in terms of design style, as shown in Figure 2. Speaking of which, we applied K-Means [11] to cluster the lab embeddings to find out what is a good number of groups for the 1,314 labs available. Our result, using the Elbow Method, is that we can comfortably put them into nine groups. Such clustering capabilities of our model could be instrumental in investigating the relationship between these labs.

We also expect this model to be very robust for unseen labs. As pointed out in Section 1, we can set a threshold and use the unseen embedding to detect that a DNA sequence comes from an unknown lab. Besides that, suppose that an accidental release happened and we need to find out the responsible. Even though some labs are not part of our training dataset, we could ask for some samples of sequences designed and extract their embeddings. We could average these embeddings to generate an embedding for the lab. With this embedding at hand, we could compute this lab's similarity with the DNA sequence that was accidentally released. It is worth noting that we can do that without requiring the model to be retrained.

For last, these embeddings are also very feature-rich, which means we can use them as input for other machine learning models, tackling other problems. For example, we know that it is common to design DNA sequences derived from multiple different sources. Suppose we want to identify the composition of a given DNA. We could prepare a dataset and use our embeddings to serve as inputs

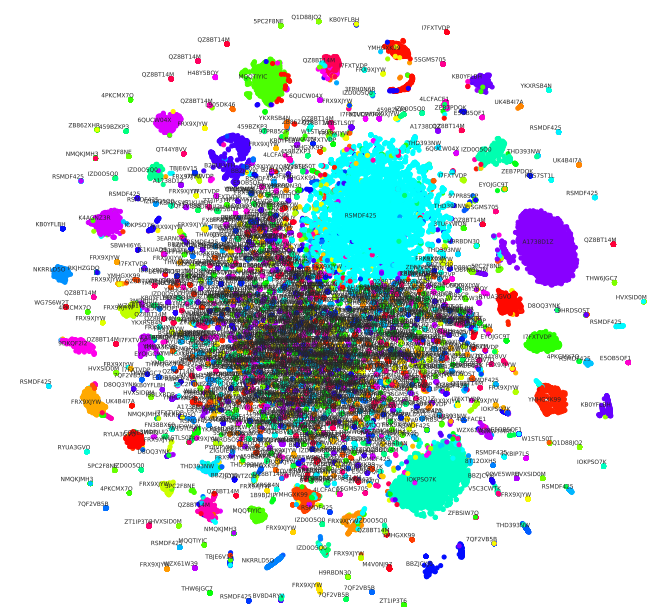


Figure 2: We used t-SNE [7] to project 200D into 2D and produce this Figure. Each circle represents a DNA sequence, and the color shows the lab-of-origin of such sequence. The lab is also presenting using a text with its ID. We can also see that there are many sequences very similar in the middle, even though they come from different labs. We theorize that they are from small labs that use similar design techniques. It is also worth noting that dimensionality reduction causes some losses, which might also explain it.

for another model. It would drastically accelerate the creation of such a model.

3 RESULTS

During the competition, we tuned the classification model and triplet network to improve their performance. Our best triplet network got [redacted]% top ten accuracy in the test dataset, while our best classification model got [redacted]%. It shows that the triplet network performs better than a classification model in such a ranking task.

Since these models give very different outputs, it was very tough to combine them in an ensemble. We first tried to compute each lab’s average rank position known as Borda voting rule [3]. However, the results were worse than the individual models. We then decided to try the Copeland voting rule [13], which gave us a much better result. Our final submission using this ensemble got [redacted]% top ten accuracy in the test dataset and the [redacted] position in the competition.

4 CONCLUSIONS AND FUTURE WORK

Metric Learning is quite common in other areas (like Face Recognition and Recommender Systems). However, for the best of our

knowledge, it has never been used in this context. We showed it is an innovative solution with promising results, outperforming a classification model trained with Supervised Learning. Moreover, it can also generate embeddings as a by-product, which are very useful.

We believe that more advanced techniques could be applied to extract features from the DNA sequences to improve it even further. New techniques like Transformers and Graph Convolutional Networks would be good candidates for this task. With better feature extraction, the embeddings would have higher quality, and the overall accuracy would also improve.

REFERENCES

- [1] [n.d.]. Genetic Engineering Attribution Challenge . <https://www.drivendata.org/competitions/63/genetic-engineering-attribution/page/164/>. <https://www.drivendata.org/competitions/63/genetic-engineering-attribution/page/164/> Accessed: 2020-10-22.
- [2] Ethan C. Alley, Miles Turpin, Andrew Bo Liu, Taylor Kulp-McDowall, Jacob Swett, Rey Edison, Stephen E. Von Stetina, George M. Church, and Kevin M. Esvelt. 2020. Attribution of genetic engineering: A practical and accurate machine-learning toolkit for biosecurity. *bioRxiv* (2020). <https://doi.org/10.1101/2020.08.22.262576> arXiv:<https://www.biorxiv.org/content/early/2020/08/22/2020.08.22.262576.full.pdf>
- [3] J-C de Borda. 1781. Mémoire sur les élections au scrutin: Histoire de l’Académie Royale des Sciences. *Paris, France* 12 (1781).
- [4] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2020. Robust Out-of-distribution Detection for Neural Networks. arXiv:2003.09711 [cs.LG]
- [5] Philip Gage. 1994. A New Algorithm for Data Compression. *C Users J.* 12, 2 (Feb. 1994), 23–38.
- [6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. arXiv:1703.07737 [cs.CV]
- [7] Geoffrey Hinton and Sam Roweis. 2002. Stochastic Neighbor Embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS’02)*. MIT Press, Cambridge, MA, USA, 857–864.
- [8] Elad Hoffer and Nir Ailon. 2018. Deep metric learning using Triplet network. arXiv:1412.6622 [cs.LG]
- [9] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882 [cs.CL]
- [10] Brian Kulis. 2013. Metric Learning: A Survey. *Foundations and Trends® in Machine Learning* 5, 4 (2013), 287–364. <https://doi.org/10.1561/22000000019>
- [11] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- [12] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. 2019. Likelihood Ratios for Out-of-Distribution Detection. arXiv:1906.02845 [stat.ML]
- [13] Donald G. Saari and Vincent R. Merlin. 1996. The Copeland method. *Economic Theory* 8, 1 (01 Feb 1996), 51–76. <https://doi.org/10.1007/BF01212012>