Deciphering Lifespan Variability in Mammalian Genomes with Predictive Modeling

Wyatt McCarthy and Niall Murphy

Spring 2024

Overview

The hypothesis behind our research comes from David Sabitini, who suspects that some fundamental component of DNA may influence a species' maximum lifespan. David speculates that this phenomenon is analogous to the longevity of cars, which is often determined by a specific or fundamental component of the vehicle's design and engineering.

To explore this hypothesis, we are investigating evolutionarily conserved mammalian gene signatures across 453 mammalian species. The nature of *conserved* gene signatures suggests an importance that has withstood evolution. Therefore, features determining fundamental aspects of fitness, such as maximum longevity, may be more prominent in the DNA of *conserved genes*.

Orthologs are a type of homologous, conserved gene that stem from a speciation event and retain similar functions.[1] In essence, an orthologous gene between humans and another species, detonated as '*s'*, indicates that the gene has an identical function and similar DNA composition in both humans and *'s'*.

Leveraging an open-source dataset[2] containing approximately 24 million DNA sequences orthologous to the human genome, our research focused on organizing, cleaning, and statistically analyzing the provided data. Additionally, we compiled and extracted various supplementary data that could contribute to our investigation. Our analysis specifically aimed to identify subsets of similar DNA sequences for which there was a statistically significant difference between the subset's maximum lifespans and the parent set's maximum lifespans. The findings from this statistical analysis provided valuable insights that have guided our model's development and training so far.

Lifespan Data

We have max lifespan data for 344 of the 453 total species. We collected our max lifespan data primarily (305/344 entries) from the AnAge database[3], a component of the HAGR (Human Ageing Genomic Resources) databases. AnAge has been described as the 'gold standard' of animal longevity data given its manual curation, thorough quality assurance, and consistent maintenance. Our other source for lifespan

---

[1] Libretexts. (2022, December 24). *7.13c: Homologs, orthologs, and paralogs*. Biology LibreTexts. https://bio.libretexts.org/Bookshelves/Microbiology/Microbiology_(Boundless)/07%3A_Microbial_Genetics/7.13%3A_Bioinformatics/7.13C%3A_Homologs_Orthologs_and_Paralogs

[2] Bogdan M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale. *Science***380**,eabn3107(2023).DOI:10.1126/science.abn3107

[3] Tacutu, R. *et al*. Human Ageing Genomic Resources: new and updated databases. *Nucleic acids research* 46, D1083–d1090, https://doi.org/10.1093/nar/gkx1042 (2018).

data (39/344 entries) was AnimalDiversityWeb, a database maintained by the University of Michigan Museum of Zoology.[4]

DNA sequence data pertaining to species for whom we could not find max lifespan data was excluded from the analysis.

DNA Sequence Data

The Zoonomia project has collected mass quantities of human-orthologous gene data through TOGA (Tool to infer Orthologs from Genome Alignments), a machine-learning classification software that detects and annotates orthologous gene loci with very high accuracy. The comprehensive data is available here.[5]

We've extracted high-quality ortholog data from TOGA's collection; for each extracted sequence, we compiled the following information from TOGA's data and annotations:

- DNA Sequence
- Chromosome
- Start/End coordinate (in standard genomic coordinates)
- Orthology Type
- Intactness
- Organism/Species
- Genome Accession

Orthology Type

Orthologs are categorized into four ortholog types (1:1, 1:Many, Many:1, Many:Many). Eventually, we want to fully understand the implications of ortholog type on data's validity so we can account for it in our data quality control.

Intactness

TOGA records another metric, intactness, to classify the likelihood that an orthologous gene encodes a functional protein. Intactness measurements are classified as follows:

---

[4] Myers, P., R. Espinosa, C. S. Parr, T. Jones, G. S. Hammond, and T. A. Dewey. 2024. The Animal Diversity Web (online). Accessed at https://animaldiversity.org.

[5] Bogdan M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale. *Science***380**,eabn3107(2023).DOI:10.1126/science.abn3107

(i) “intact”: the middle 80% of the CDS (coding sequence) is present and exhibits no gene-inactivating mutation; likely to encode functional proteins

(ii) “partially intact” : ≥50% of the CDS is present and the middle 80% of the CDS exhibits no inactivating mutations; may also encode functional proteins but the evidence is weaker

(iii) “missing”: <50% of the CDS is present and the middle 80% of the CDS exhibits no inactivating mutation; undecided protein encoding because more than half of the CDS is missing but no strong evidence for loss exists

(iv) “uncertain loss”: at least one inactivating mutation in the middle 80% of the CDS, but evidence is not strong enough to classify the transcript as lost; may or may not encode a functional protein;

(v) “lost”: evidence for loss is sufficiently strong; unlikely to encode a functional protein.

For our analysis, we excluded all sequences annotated as ‘missing’, ‘uncertain loss’, or ‘lost’. We also excluded all sequences where over 25% of the base pairs consisted of gaps (denoted as ‘-’ in a sequence). In the future, we are considering constraining analysis based on orthology type but need to better understand the implications of each type before doing so.

Identifying DNA of Interest

In the early stages of my research, we were considering various components of DNA on which to narrow our focus. Prior work on the project and suggestions from David pointed to certain *consensus sequences* and *transcription factor binding sites* as key areas of interest.

Consensus Sequences

Consensus sequences are found within conserved gene signatures and defined as: a theoretical representative nucleotide or amino acid sequence in which each nucleotide or amino acid is the one which occurs most frequently at that site.[6] *Consensus sequences* (by nature) do not occur out of coincidence and hold important functions within the conserved signature in which they’re found. Examples of such sequences are the TATA box and Kozak Sequence, which occur in the promoter region of DNA.

[6] U.S. National Library of Medicine. (n.d.). *Consensus sequence*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Consensus%2BSequence%22%5BMeSH%2BTerms%5D

Transcription Factor (TF)

Transcription factors are the primary regulators of gene expression; they are regulatory proteins that bind to specific DNA sequences (transcription factor binding sites) and regulate when, where, and the extent to which genes are expressed.[7]

Transcription Factor Binding Sites (TFBS)

Transcription factor binding sites refer to regions of DNA, usually 6-8 bases long and upstream of the transcription start site, that transcription factors (regulatory proteins) bind to as a prerequisite step for the recruitment of RNA polymerase and eventual transcription. Given their role in transcription regulation, these binding sites are of pivotal importance in the eventual expression of a gene.

- Example of significance:
  - "Interestingly, a single mutation in one of the zinc finger motifs of the kruppel protein which replaces a cysteine by a serine that cannot bind zinc, results in a mutant fly whose appearance is exactly identical to that produced by complete deletion of the gene".[8]

Due to their influence on gene expression, analyzing the variance in certain TFBS across many mammalian species may yield novel insights into the relationship between binding sites, their mutations, corresponding transcription factors, and maximum lifespan.

The orthologs provided by the Zoonomia project contained only *coding region* DNA. However, to analyze the suggested consensus sequences (TATA box and Kozak sequence) and TFBS, we needed to extract the *promoter region* DNA corresponding to each entry in our data.

Coding Region

The *coding region* describes the region of DNA downstream from the transcription start site, that is transcribed into mRNA.

Promoter Region

---

[7] Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P., & Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (Review). World Academy of Sciences Journal, 2, 3-8. https://doi.org/10.3892/wasj.2020.32

[8] Latchman DS. Eukaryotic transcription factors. Biochem J. 1990 Sep 1;270(2):281-9. doi: 10.1042/bj2700281. PMID: 2119171; PMCID: PMC1131717.

> The *promoter region* describes the region of DNA upstream from the transcription start site; the promoter region is where transcription factors (regulatory proteins) bind to initiate the transcription of a gene. Transcription factor binding sites are found in the promoter region.

Using TOGA's orthologous gene data, we extracted additional DNA sequences from UCSC[9] and DNA Zoo Consortium[10] genome assemblies. More formally, given an orthologous sequence, its genome accession, the chromosome in which it is located, and its transcript start/end coordinates by TOGA, we extracted the DNA sequences that fall 1000 base pairs up/downstream (depending on sequence orientation) of the transcription start site from corresponding genome assemblies made available by UCSC and DNA Zoo Consortium.

As known consensus sequences like the TATA box and Kozak sequence are found proximal to the transcription start site, having access to regions of DNA 1000 base pairs up/downstream from a conserved sequence's transcription start site provides a valuable window from which to determine the existence of consensus sequences and other TFBS in the promoter region such that we can analyze their features.

We planned to perform such analysis primarily using BioPython.[11] With the functionality provided by BioPython, we could search promoter region DNA for instances of each consensus sequence and record relevant features. We've mainly considered the following key features:

- The variant of the consensus sequence that appears (e.g. the TATA box consensus sequence is TATAWAWN, where W=A/T and N=A/T/G/C. Thus, many TATA box consensus sequence variants can appear.)
- The location of the identified consensus sequence relative to the transcription start site

However, due to limitations in the annotations provided by TOGA, we were only able to extract promoter region DNA for approximately 8 million of the total 24 million sequences. This significant reduction in sample size led us to shift our focus to analyzing coding regions of DNA, as provided out-of-the-box by the Zoonomia Project.

---

[9] Nassar et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research* 2023 PMID: 36420891, DOI: 10.1093/nar/gkac1072

[10] Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds." Science (New York, N.Y.) 356 (6333): 92-95. https://doi.org/10.1126/science.aal3327

[11] Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009 Jun 1; 25(11) 1422-3 https://doi.org/10.1093/bioinformatics/btp163 pmid:19304878

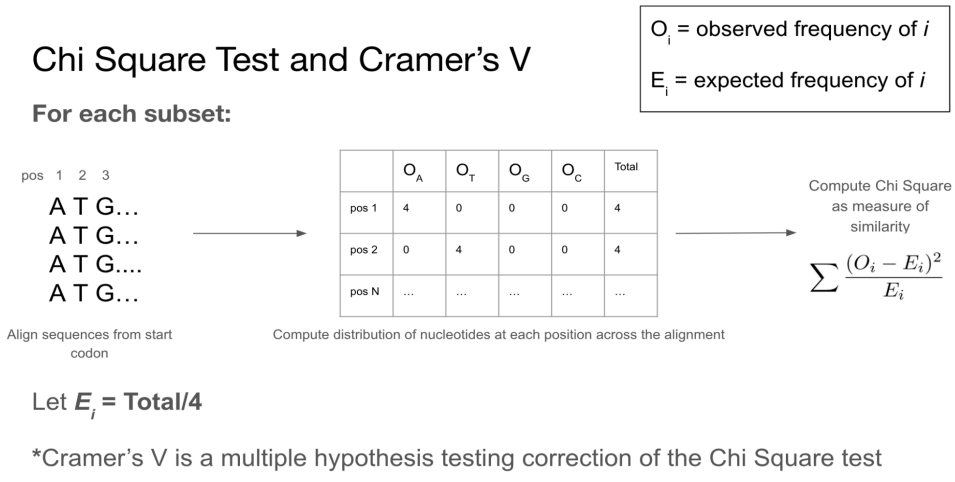
If we revisit consensus sequence analysis in the future, we aim to identify a more comprehensive set of features that may influence the role of each consensus sequence. We would also like to compile an extensive list of consensus sequences of interest, such that our analysis is not constrained to just the TATA box and Kozak sequence.

For the time being, we've narrowed our focus to orthologs that code for transcription factors (TFs) due to the paramount role of TFs in gene expression. Cross-referencing a TF database[12] of known human/mouse/rat TFs with TOGA's gene annotations, we organized all orthologs into subsets categorized by the gene/TF that they code for. Our analysis to date has focused on these TF ortholog sets.

The Data Pipeline

Having organized all orthologs into sets according to the transcription factor that they code for, we ran each set through the following pipeline:

1. **Gather a preliminary idea of sequence similarity across set**
    a. Align sequences in set from the start codon and compute the distribution of nucleotides at each position of sequence alignment
    b. Compute chi square based on the distribution of each nucleotide in each position
    c. Compute cramer's V (multiple hypothesis testing correction) from chi square
    d. Chi Square and Cramer's V allowed us to glean a proxy for sequence similarity within a set; that is, the higher the Cramer's V/Chi Square for a set, the more similarity exists across its sequences and vice-versa



	O_A	O_T	O_G	O_C	Total
pos 1	4	0	0	0	4
pos 2	0	4	0	0	4
pos N	...	...	...	...	...

[12] Kuiper, M., & Lægreid, A. (n.d.). *Transcription Factor Checkpoint 2.0*. TFCheckpoint. https://tfcheckpoint.org/index.php

2. **Embed all sequences in set**
    a. Utilized **DNABERT-S**[13] to create species-aware embeddings

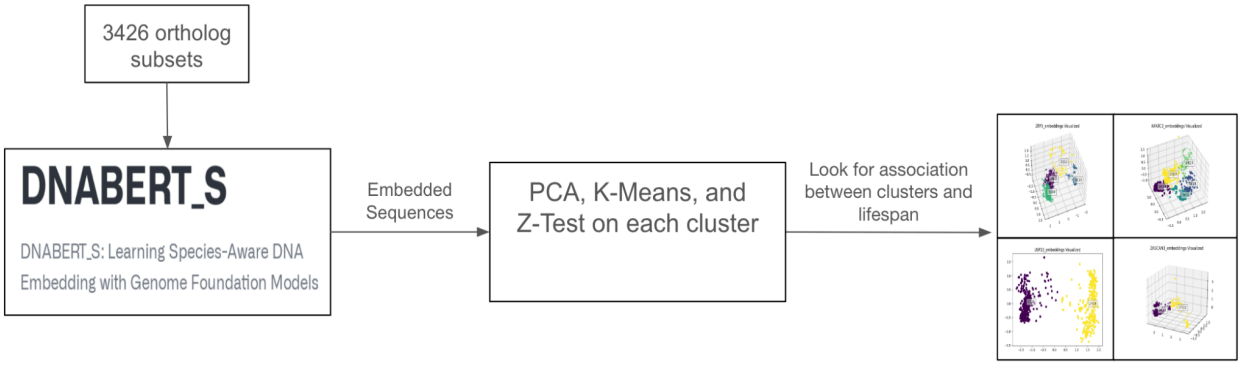
3. **Principle Component Analysis and K-means clustering on a set's embeddings**
    a. Utilized sci-kit learn, scipy for computation, and matplotlib to visualize embedding space
    b. In visualized embedding space, **certain sets display clusters that appear to be associated with lifespan differences** (see below)

4. **For each cluster in a set, compute modified z score and statistical significance (p value)**
    a. Given how our modified z score was defined, the more statistically significant a cluster, the higher the correlation between lifespan and DNA composition specific to that cluster
    b. The collected stats help to reinforce the findings of our visual analysis on embeddings

$$Modified\ Z\ Score\ =\ \frac{.6745\ (M_{cluster} - M_{set})}{MAD_{set}}$$

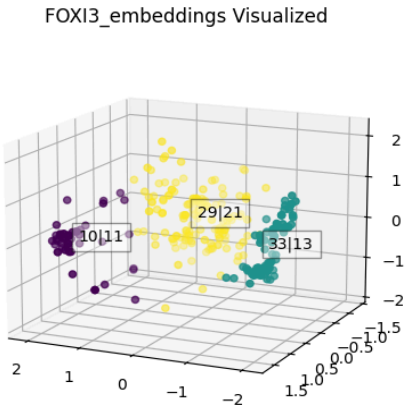
$$M_{cluster}\ =\ Median\ max\ lifespan\ of\ cluster$$
$$M_{set}\ =\ Median\ max\ lifespan\ of\ parent\ set$$
$$MAD_{set}\ =\ Median\ absolute\ deviation\ of\ parent\ set$$

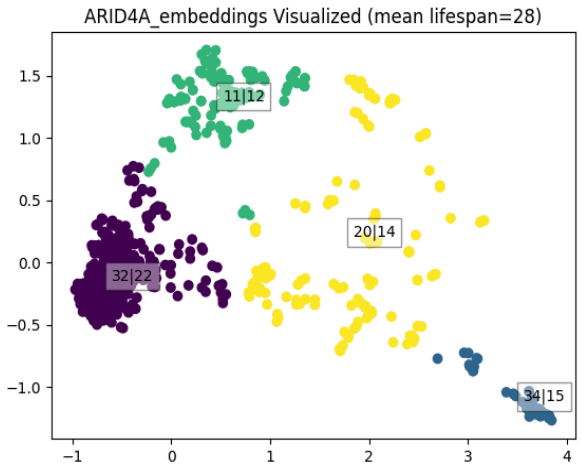
Steps 2-4



[13] Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., … Liu, H. (2024). DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models. *arXiv [q-Bio.GN]*. Retrieved from http://arxiv.org/abs/2402.08777

Embedding Visualizations on Sets with High Z Scoring Clusters





Findings Inform our Modeling Approach

Though our statistical analysis helps to identify ortholog sets that exhibit some correlation between lifespan and sequence variation, the data is not nearly robust enough to conclude causation. Nevertheless, the findings inform our modeling approach.

A straightforward modeling approach for our data would involve training a model to predict maximum lifespan given a DNA sequence. With that being said, the trends we've noticed throughout our analysis lead us to believe that this is an impossible task if not properly defined. By that, there doesn't seem to be a DNA feature that is universally predictive of lifespan; in other words, we could not train a model to predict max lifespan given an *arbitrary* DNA sequence. Instead, it may be possible to train a model to predict lifespan given DNA from specific genes or gene subsets, where we have observed correlations between lifespan and DNA variation.

In any case, we need training data with a clear and learnable association between lifespan and DNA composition variance. As such, our statistical analysis proves valuable in identifying which sets across our cumulative data may be useful in training.

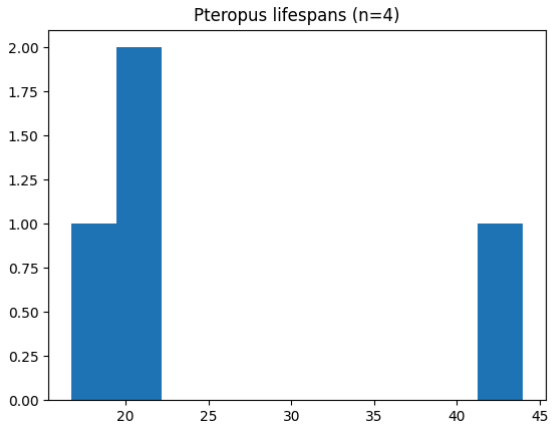
Specifically, the z-test we performed allows us to eliminate sets with low variance and/or low z-scoring clusters from consideration for training data; if there is no association between lifespan and sequence variance within a set, there is nothing useful our model can learn from its DNA sequences. Conversely,

we can identify high z-scoring sets with statistically significant clusters where the maximum lifespan differs from the parent set. These ‘good’ sets may make for more informative and optimal training data.

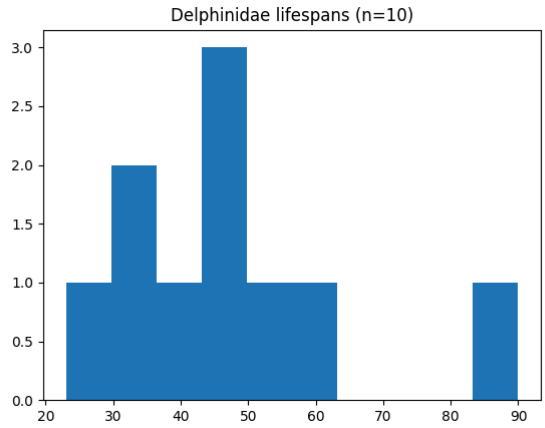
Regardless, even if our model can learn from this ‘good’ data, the main challenge arises in validating that the DNA features the model learns to associate with lifespan variance are not simply the product of randomness or noise. As such, we reasoned that the model could be evaluated based on its ability to accurately predict lifespan given DNA from an *‘outlying species’*. We defined *‘outlying species’* as a species whose max lifespan is an outlier among all other species from the same taxonomic classification; the lower the classification on the taxonomic hierarchy, the better.

To identify outlying species’, we organized our max lifespan data by taxonomic genus (lowest classification), family (2nd lowest), and order (3rd lowest) then performed simple outlier detection. A sample of taxonomic groups in which we identified ‘outlying species’ are plotted below:

Pteropus Genus



Delphinidae Family



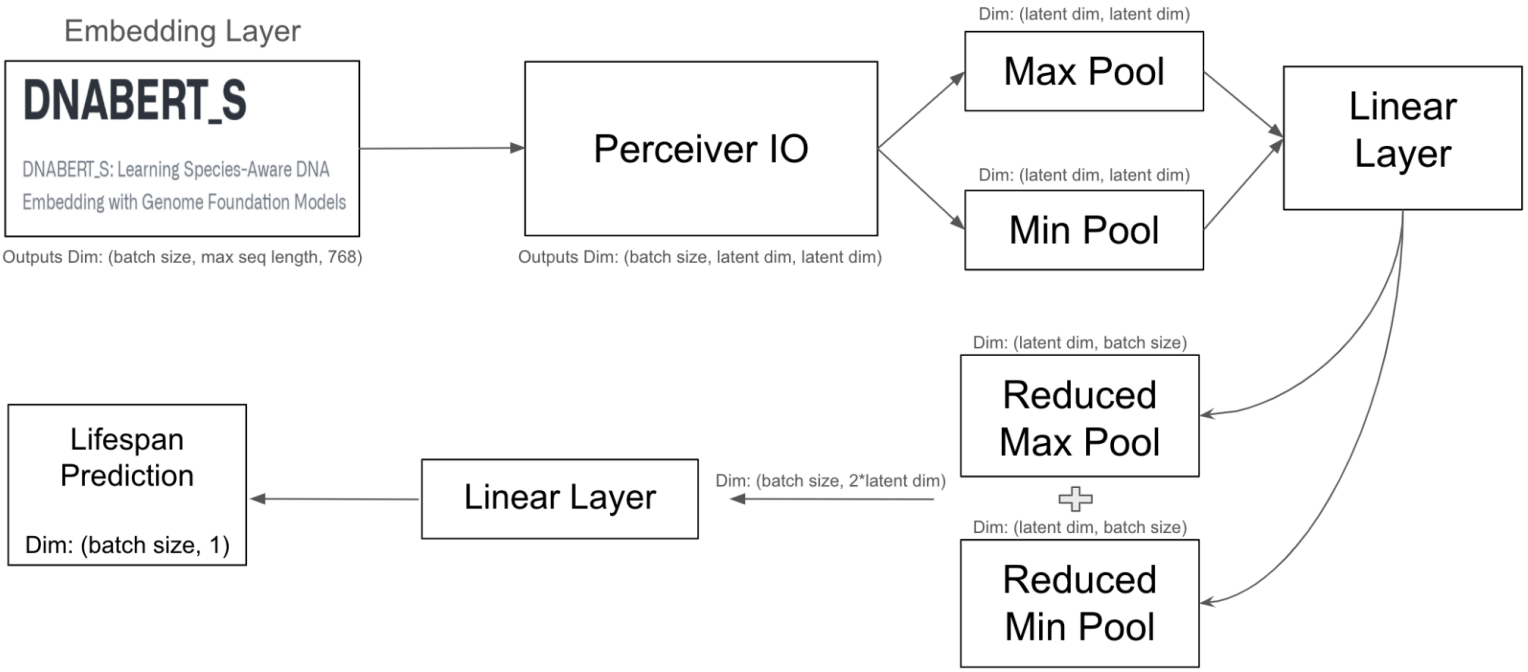
We have not yet trained the model to a point where it can be evaluated as described, but the ‘outlying species’ benchmark is certainly a promising avenue.

Model Architecture

Given our data's high maximum sequence length, a Perceiver-based architecture made the most sense to avoid the quadratic scaling usually had in Transformer self-attention.

As of now, our model consists of the following layers (in order):

- Embedding Layer (DNABERT-S)
    - Output Dimensionality = (batch size, max length, 768)
- PerceiverIO[14]
    - Output Dimensionality = (batch size, latent dim, latent dim)
- Max + Mean Pool
    - Each pooling has dimensionality = (latent dim, latent dim)
- Linear Layer 1
    - Put max and mean pooling through linear layer
    - Output for each pooling dimensionality = (latent dim, batch size)
- Concatenate and Transpose Pooling
    - Concatenate and transpose mean and max pool
    - Output Dimensionality = (batch size, 2*latent dim)
- Linear Layer 2 / Output
    - Output Dimensionality = (batch size, 1)

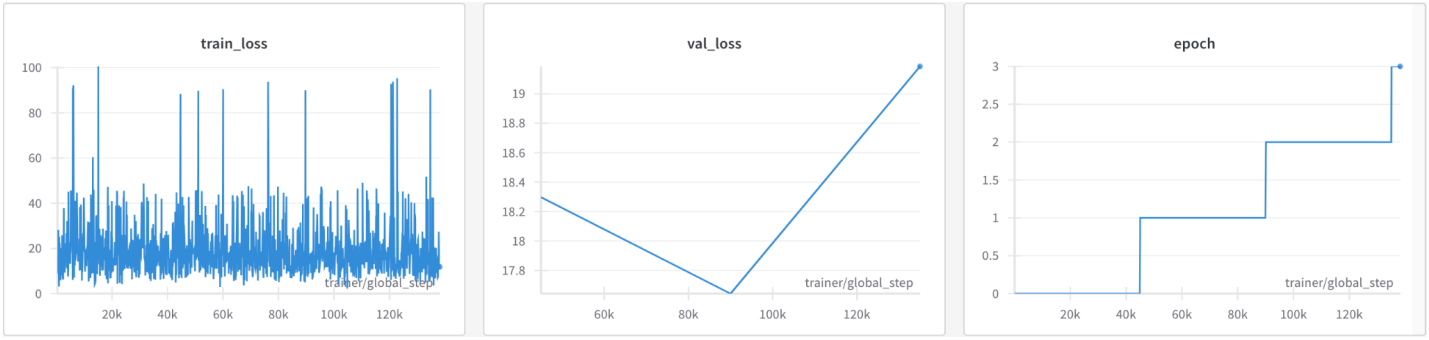


[14] Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., … Carreira, J. (2021). Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2107.14795

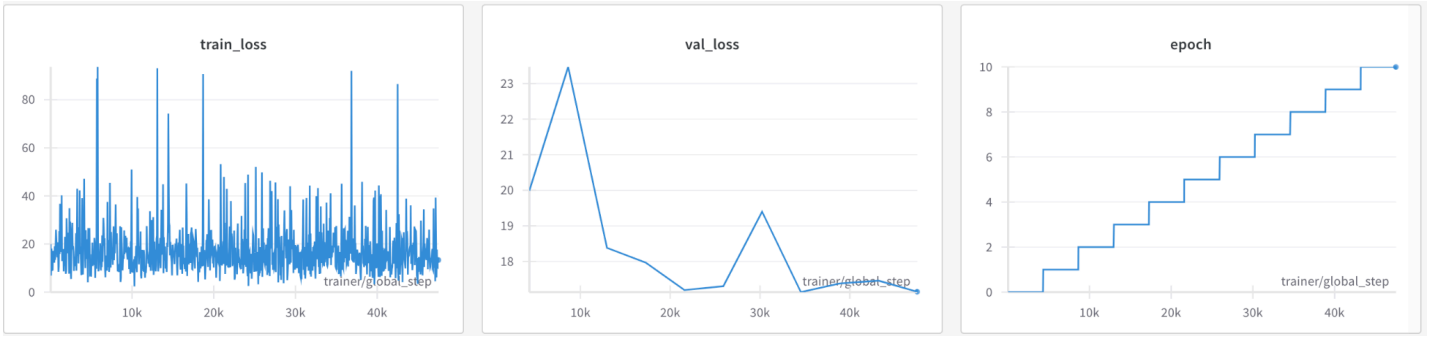
Early Modeling Results

We've only progressed to the early stages of training, focusing on evaluating how the model learns differently on arbitrarily selected DNA training data compared to training DNA from statistically significant, 'good', data subsets. The results are as follows:

Arbitrarily Selected Training DNA



Top 50 "Good" DNA Subsets



The model does not appear to learn well in either instance. This primarily suggests that we need to:

- Evaluate what constitutes 'good' training data
- Continue statistical analysis to better understand data
- Pre-train DNABERT-S embeddings
- Rethink modeling approach

Potential Next Steps

We are in the process of evaluating several options that present themselves as promising avenues for further exploration.

One potential path involves conducting additional modeling experiments to evaluate how the model's learning capabilities change when trained on arbitrarily selected DNA training data compared to training data derived from statistically significant data subsets.

On the other hand, we may scrap the PerceiverIO approach given our model's poor learning thus far, and experiment with different long-context modeling architectures such as Hyena[15] and Mamba[16]. These architectures have demonstrated strong and/or state-of-the-art performance in handling long biological sequences and may be more suitable for our problem than a PerceiverIO-based architecture.[17]

Multiple sequence alignment (MSA) techniques may also prove beneficial in capturing the evolutionary relationships and patterns within our DNA sequence data. MSA techniques take "into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions".[18] Demonstrated by MSA's successful application in state-of-the-art systems like AlphaFold, it enables deep learning models to learn more robust and interpretable representations of biological sequences.[19]

It would also be valuable to run the protein sequence counterparts of transcription factor (TF) DNA through the same data pipeline. Confirming that the DNA subsets found to be statistically significant align with the protein subsets found to be statistically significant would reinforce our findings thus far. Additionally, revisiting other collected data may provide additional insights. Particularly, we are

---

[15] Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., … Ré, C. (2023). Hyena Hierarchy: Towards Larger Convolutional Language Models. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2302.10866

[16] Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2312.00752

[17] Nguyen, E., Poli, M., & Faizi, M. (2023a, June 29). *HyenaDNA: Learning from DNA with 1 million token context*. Hazy Research. https://hazyresearch.stanford.edu/blog/2023-06-29-hyena-dna

[18] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, Cedric Notredame, Multiple sequence alignment modeling: methods and applications, *Briefings in Bioinformatics*, Volume 17, Issue 6, November 2016, Pages 1009–1023, https://doi.org/10.1093/bib/bbv099

[19] AF2 is here: what's behind the structure prediction miracle. *Oxford Protein Informatics Group*. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/.

considering running the non-TF ortholog sets through the data pipeline, or returning to consensus sequence analysis on our promoter region DNA.

Data

Zoonomia Project

TOGA

DNA Zoo Assemblies

TF Checkpoint

Our Project's GitHub

References

AF2 is here: what's behind the structure prediction miracle. *Oxford Protein Informatics Group*. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/.

Bogdan M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale. *Science***380**,eabn3107(2023).DOI:10.1126/science.abn3107

Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009 Jun 1; 25(11) 1422-3 https://doi.org/10.1093/bioinformatics/btp163 pmid:19304878

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds." Science (New York, N.Y.) 356 (6333): 92-95. https://doi.org/10.1126/science.aal3327

Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2312.00752

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., … Carreira, J. (2021). Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2107.14795

Kuiper, M., & Lægreid, A. (n.d.). *Transcription Factor Checkpoint 2.0*. TFCheckpoint. https://tfcheckpoint.org/index.php

Latchman DS. Eukaryotic transcription factors. Biochem J. 1990 Sep 1;270(2):281-9. doi: 10.1042/bj2700281. PMID: 2119171; PMCID: PMC1131717.

Libretexts. (2022, December 24). *7.13c: Homologs, orthologs, and paralogs*. Biology LibreTexts. https://bio.libretexts.org/Bookshelves/Microbiology/Microbiology_(Boundless)/07%3A_Microbial_Genetics/7.13%3A_Bioinformatics/7.13C%3A_Homologs_Orthologs_and_Paralogs

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, Cedric Notredame, Multiple sequence alignment modeling: methods and applications, *Briefings in Bioinformatics*, Volume 17, Issue 6, November 2016, Pages 1009–1023, https://doi.org/10.1093/bib/bbv099

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P., & Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (Review). World Academy of Sciences Journal, 2, 3-8. https://doi.org/10.3892/wasj.2020.32

Myers, P., R. Espinosa, C. S. Parr, T. Jones, G. S. Hammond, and T. A. Dewey. 2024. The Animal Diversity Web (online). Accessed at https://animaldiversity.org.

Nassar et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research* 2023 PMID: 36420891, DOI: 10.1093/nar/gkac1072

Nguyen, E., Poli, M., & Faizi, M. (2023a, June 29). *HyenaDNA: Learning from DNA with 1 million token context*. Hazy Research. https://hazyresearch.stanford.edu/blog/2023-06-29-hyena-dna

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., … Ré, C. (2023). Hyena Hierarchy: Towards Larger Convolutional Language Models. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2302.10866

Tacutu, R. *et al*. Human Ageing Genomic Resources: new and updated databases. *Nucleic acids research* 46, D1083–d1090, https://doi.org/10.1093/nar/gkx1042 (2018).

U.S. National Library of Medicine. (n.d.). *Consensus sequence*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Consensus%2BSequence%22%5BMeSH%2BTerms%5D

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., … Liu, H. (2024). DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models. *arXiv [q-Bio.GN]*. Retrieved from http://arxiv.org/abs/2402.08777