Species-aware DNA language modeling

Dennis Gankin*[1], Alexander Karollus*[1], Martin Grosshauser[1], Kristian Klemon[1], Johannes Hingerl[1], Julien Gagneur[1,2,3,4,5]

[1]School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[2]Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
[3]Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany
[4]Munich Data Science Institute, Technical University of Munich, Garching, Germany
[5]Munich Center for Machine Learning, Germany
*Contributed equally
Correspondence to: gagneur@in.tum.de

Abstract

Motivation

Predicting gene expression from DNA is an open field of research. As in many areas, labeled data is dwarfed by unlabelled data, i.e. species with a sequenced genome but no gene expression assay data. Pretraining on unlabelled data using masked language modeling has proven highly successful in overcoming data constraints in natural language and proteomics. However, in genomics, this approach has so far been applied only to single genomes, neither leveraging conservation of regulatory sequences across species nor the vast amount of available genomes.

Results

Here we train a masked language model on more than 800 species spanning over 500 million years of evolution. We show that explicitly modeling species is instrumental in capturing conserved yet evolving regulatory elements and in controlling for oligomer biases. We extract embeddings for 3' untranslated regions of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and use them to achieve prediction of mRNA half-life that is better or on-par with the state-of-the-art, demonstrating the utility of the approach for regulatory genomics. Moreover, we show that the per-base reconstruction probability of our model significantly predicts RNA-binding protein bound sites directly. Altogether, our work establishes a self-supervised framework to leverage large genome collections of evolutionary distant species for regulatory genomics and contributes to alignment-free comparative genomics.

Availability and implementation

The source code and trained models are available at:
https://github.com/DennisGankin/species-aware-DNA-LM

Introduction

A typical eukaryotic genome contains large regions of non-coding DNA. These are not translated into proteins, but contain regulatory elements which control gene

expression in response to environmental cues. Finding these regulatory elements and how their combinations and arrangements determine gene expression is a major goal of genomics research and is of great utility for synthetic biology. Because of the high complexity of the underlying regulatory code, supervised machine learning approaches have become established as a key computational method to tackle this task (Alipanahi *et al.*, 2015; Zhou and Troyanskaya, 2015; Eraslan *et al.*, 2019).

However, an Achilles heel of supervised approaches in regulatory genomics is that their training requires experimentally-derived labels. Many of these experiments require considerable resources and expertise, thus limiting the amount of data available.

A key insight in biology is that the genomes of all organisms are evolutionarily related. This insight can be leveraged to study regulatory elements through sequence comparison without labels (Kellis *et al.*, 2003). Specifically, we expect that regulatory sequences and functional arrangements thereof will be selected and thus generally more conserved than expected under neutral mutation (Kimura, 1968). The main method to determine whether particular nucleotides are conserved makes use of sequence alignment. Unfortunately, alignment is difficult for non-coding sequences, presumably because regulatory sequences evolve faster than coding sequences and because of a certain tolerance to the exact order, orientation, and spacing of regulatory elements in regulatory sequences.

Masked language modeling (MLM), a type of self-supervised learning developed for natural language processing, presents another way to generate insights from unlabelled data (Devlin *et al.*, 2019). In MLM, parts of an input sequence are hidden (masked) and a model is tasked to reconstruct them. Models trained in this way learn syntax and semantics of natural language and achieve state-of-the-art performance on many downstream tasks.

Previous work has established MLM, specifically masked *nucleotide* modeling, as a method to pre-train models for genomics (Ji *et al.*, 2021; Benegas *et al.*, 2022). It was shown that these models capture some semantics of the regulatory code. However, this work generally focused on individual species and included both coding and non-coding regions. These models do not benefit from evolutionary relationships except for in-genome homology. Moreover, as they are species-agnostic, they cannot account for the evolution of the regulatory code itself even if trained on more than one genome.

We hypothesized that MLM should be able to leverage evolutionary information without requiring alignment. Specifically, we expect that nucleotides with regulatory meaning should be more conserved and therefore easier to reconstruct when masked. Thus, selection should guide a model trained using MLM towards learning the regulatory code. To test this, we take an orthogonal approach to previous work. Specifically, we focus solely on one particular type of non-coding region, namely sequences 3‘ to stop codons, but in a vast multispecies dataset, comprising 806 fungal species, some separated by more than 500 million of years of evolution. We focus on this particular region because it is known to be important for regulating mRNA half-life, a central determinant of gene expression that is traditionally measured using the expensive and difficult to scale metabolic labeling technique.

We train a deep model on these sequences and then evaluate the ability of our model to reconstruct DNA and predict protein binding and mRNA half-life in the held-out genus Saccharomyces. To account for evolution of the regulatory code itself, we additionally provide a species label. In this way, we make the model

species-aware and can evaluate whether it also learns more specific regulatory elements.

Materials and Methods

Genome data. We obtained 1,500 fungal genomes, comprising 806 different species, from the Ensembl fungi 53 database. For each annotated protein-coding gene in each genome, we extracted 300 base pairs 3' to the stop codon of genes. While the actual transcribed 3' untranslated regions (3' UTR) vary in length, we expect that in most cases 300 bp will be sufficient to include the entire 3' UTR (Pelechano *et al.*, 2013). Overall our train set included on the order of 13 million sequences (3.9 billion nucleotides), which is similar to the size of the human genome (3.2 billion nucleotides).

As a test set, we used the annotated 3' UTR sequences (Pelechano *et al.*, 2013) of the widely studied species *Saccharomyces cerevisiae*. Note that these have variable lengths. To prevent data leakage from closely related species, the train set excludes the entire Saccharomyces genus.

Sequence alignment. For investigation purposes, sequence alignment of annotated Saccharomyces cerevisiae CDS and 3' UTR sequences was performed using discontiguous megablast, which was specifically designed for cross-species alignment, version 2.13 with default parameters (Camacho *et al.*, 2009).

Masked nucleotide modeling. We performed masked nucleotide modeling on the fungi 3' regions. Specifically, we randomly masked 15% of nucleotides in each sequence and trained a deep learning model to reconstruct these from context, so as to minimize cross-entropy. 80% of the nucleotides selected for masking were replaced with a masking token, 10% percent were randomly mutated, and the last 10% were left unchanged. Nucleotides were one-hot encoded before being passed to the model.

The deep learning model consisted of three modules. First a convolutional kernel and three residual layers with two convolutional kernels each transform the downstream sequence into a higher dimensional latent space. Then, the data passes four Diagonal State Space layers (Gupta *et al.*, 2022) and finally a convolutional kernel decodes the data to nucleotide probabilities. The State Space layer offers some advantages over transformer models which are more commonly used for MLM, including its linear scaling with sequence length. Overall the model only had 976 thousand parameters.

To make the model species-aware, the corresponding species label was provided as additional input. An embedding layer transforms this species label into a numeric vector which is added per base to the latent space. For purposes of comparison, we also trained a species-agnostic version of the model, which does not receive the species label. We trained the model for 50 epochs, as the loss and accuracy plateaued at that point.

For benchmarking, we also trained DNABERT (Ji *et al.*, 2021), a 110-million parameter species-agnostic transformer. We used the same hyperparameters as in the original publication, except that we doubled the batch-size for faster convergence. The substantial computational cost of training the model prevented us from exploring more variations thereof.

Evaluating the reconstruction. We computed reconstruction predictions for each position in the test sequences (annotated *S. cerevisiae* 3' UTR). For each position, we called the probability corresponding to the correct nucleotide the masked nucleotide probability. This score, which is inversely monotonically related to cross-entropy, was averaged over certain regions of interest. These included the consensus motifs of several RNA binding proteins (RBP) (Foat *et al.*, 2005; Hogan *et al.*, 2008; Cheng *et al.*, 2017) and experimentally validated RBP binding sites as determined by global photoactivatable-ribonucleoside-enhanced crosslinking and immunopurification (gPAR-CLIP) (Freeberg *et al.*, 2013). This experiment provides regions of the genome, typically 20 nt in length, which are bound by a protein. For comparison, we also computed mean masked nucleotide probabilities for random k-mers. Paired Wilcoxon tests with Benjamini-Hochberg correction were used to assess significance.

As the test species was held out from the training set, the model has not encountered its species label. Therefore, inference for *S. cerevisiae* was performed by providing the species label of *C. glabrata*, the species in the train set most closely related to *S. cerevisiae*.

Predictions based on masked nucleotide probabilities. We consider three tasks: predicting whether a region is gPAR-CLIP bound, with either non-bound regions or shuffled bound regions as negatives, and predicting whether consensus Puf3 motifs are bound by Puf3. To get non-binding regions we randomly sampled 20-nt regions from the areas of the *S. cerevisiae* 3' UTR which were not identified to be bound in the experiment, until class balance is achieved. For the task where bound and shuffled regions are distinguished, we create five dinucleotide shuffles for each binding site and remove duplicates. The surrounding 3' UTR sequence was not shuffled. For the Puf3 task, we collected all instances of the Puf3 consensus motifs in annotated 3' UTR. The positive set consisted of all instances in genes with experimental evidence of Puf3 binding (Lapointe *et al.*, 2018). The negative set comprised the rest. We then computed mean masked probabilities for all the regions of interest. PhastCons conservation scores (Siepel *et al.*, 2005) for *S. cerevisiae* were downloaded from UCSC and extracted for regions of interest.

Downstream tasks: mRNA half-life and reporter expression prediction. To generate embeddings, we successively masked each nucleotide in the input sequence and extracted the latent representation generated by the last state space layer. To normalize for sequence length, we averaged over nucleotides in the sequence.

For predicting half-life, we provided 5' UTR and coding sequence (CDS) features, as described in (Cheng *et al.*, 2017) for *S. cerevisiae* and (Eser *et al.*, 2016) for *S. pombe*, in addition to our embeddings (or competing sequence representations) to a ridge regression with the regularization parameter set by nested cross-validation. We performed 10-fold cross validation to estimate generalization performance.

We proceeded similarly for predicting reporter expression in the experiment of (Shalem *et al.*, 2015) but did not include any features besides the 3' UTR representation (as 5' UTR and CDS are constant in this experiment). However, as the reporter contains many almost-duplicate sequences, we performed grouped 10-fold cross-validation, where the groups consisted of 3' UTR sequences which showed a 5-mer spectrum cosine similarity above 0.7. This prevented trivial overfitting.

In all cases, paired Wilcoxon tests with Benjamini-Hochberg correction were used to assess statistical significance.

Results

Fungi 3' regions are too evolutionary distant for effective alignment even when regulatory elements are conserved

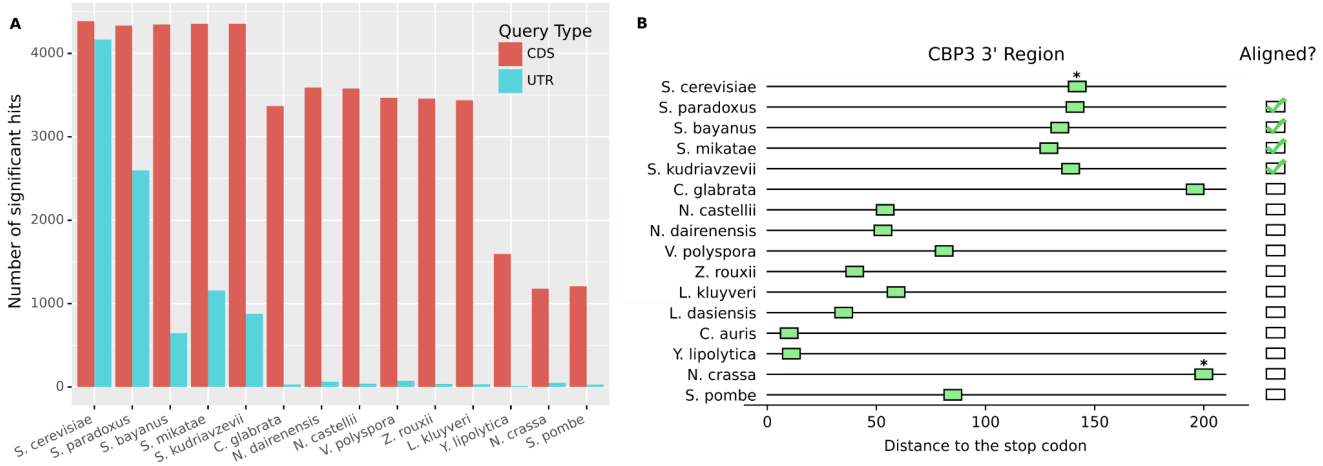

Fig. 1 Alignments struggle to capture conservation of regulatory elements in 3' regions. **A)** BLAST hits of *S.cerevisiae* CDS and 3' UTRs in other fungal species. **B)** Regions 3' to the stop codon of CBP3 orthologues in different fungal species. Instances of Puf3 like motifs (TGTA*ATA) indicated in green, star indicates experimental evidence of Puf3 binding. It appears that Puf3 binding is conserved whereas the location of the motif is not.

Sequence alignment is a well-known and highly effective method to study the evolution of biological sequences. They can be used to detect homologies, find conserved subsequences, and pinpoint sequence motifs. We first assessed whether sequence alignments could be a viable starting point to capture conserved regulatory sequence elements in 3' regions across large evolutionary distances. To this end, we aligned coding sequences and annotated 3' UTR of *Saccharomyces cerevisiae* to the genomes of a variety of fungal species (Fig. 1A). Coding sequences could be successfully aligned even between highly diverged species. In contrast, the ability to align 3' regions almost completely disappears beyond the Saccharomyces genus.

However, many regulatory sequence elements of 3' UTRs are conserved far beyond the genus boundary. The 3' region of the cytochrome B assembly factor CBP3 illustrates this. Experiments have shown that the RNA binding protein Puf3 binds 3' UTR of this gene in *S. cerevisiae* and the far diverged (~500 mio years (Prieto and Wedin, 2013)) *Neurospora crassa* (Wilinski *et al.*, 2017). Moreover, we found that the Puf3 consensus binding motif can be found 3' to the stop of the CBP3 homologue in almost all yeasts. However, we observed that the motif appears to be highly mobile, complicating alignment (Fig 1B). This example illustrates the need and potential for alignment-free approaches to leverage conservation across large evolutionary distances.

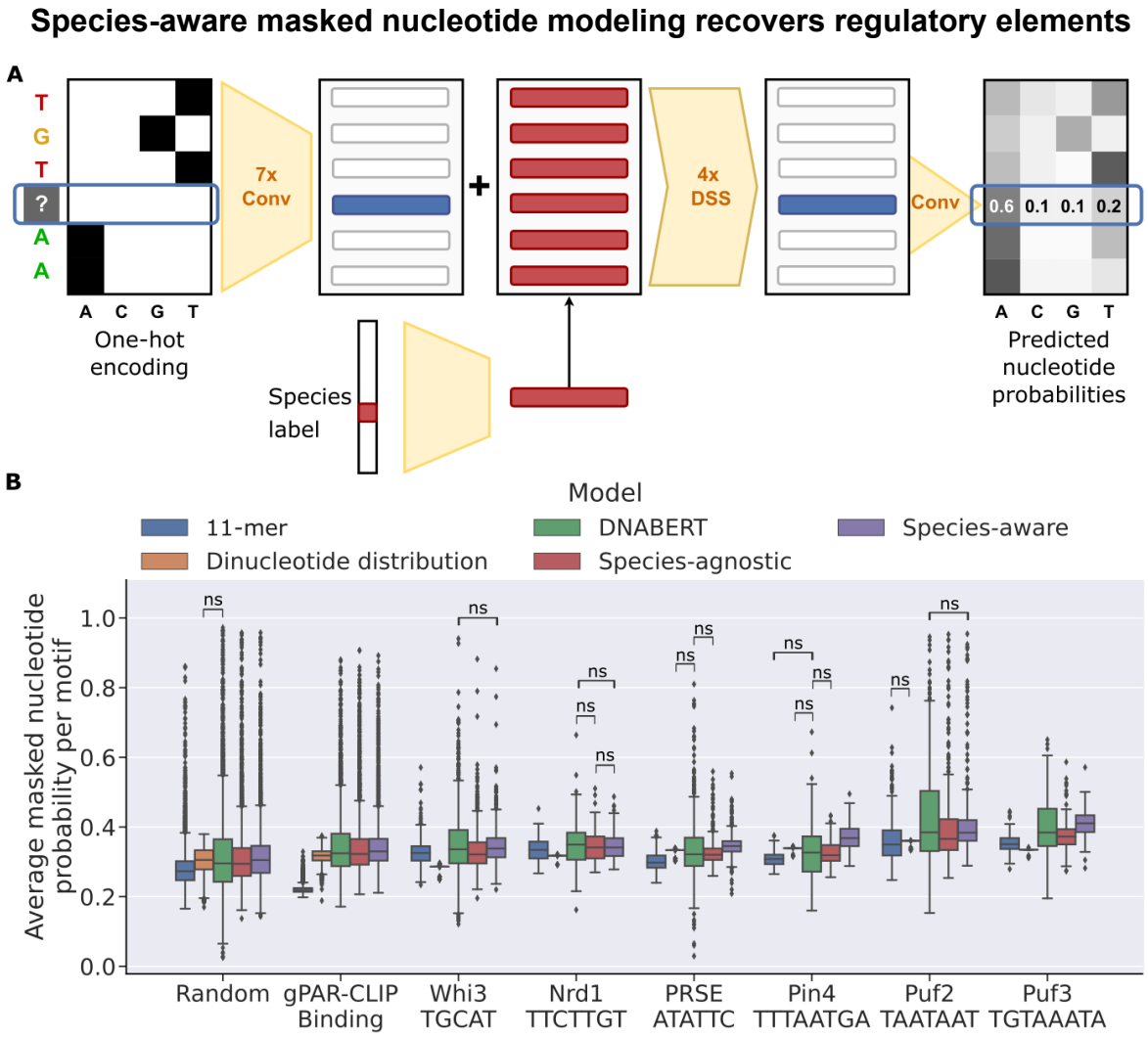

Fig. 2: Masked nucleotide modeling enables reconstruction of likely regulatory sequences in a held-out species. A) Architecture of our species-aware state space model **B)** Average masked nucleotide probabilities for instances of random 3' UTR subsequences, experimentally validated protein binding sites (gPAR-CLIP) and RNA-binding protein consensus motifs in *S. cerevisiae* 3' UTR sequences. All pairwise comparisons are statistically significant, except the ones marked not significant (ns).

We next considered alignment-free models. Because the regulatory code evolves across long evolutionary times, we designed a species-aware model, in which a species identifier is provided to the model (Fig. 2A), which we compared to a species-agnostic model with the same architecture. This common architecture is a diagonal state space model (Gupta *et al.*, 2022), a neural network that can efficiently model long-range dependencies (Gu *et al.*, 2021). We also considered a model predicting a central nucleotide according to 11-mer frequencies of the entire train dataset. The general assumption is that these models, which learn sequence distribution with various degrees of sophistication, will indirectly learn to identify regulatory sequences, as these should be statistically enriched due to selection. As baseline we fitted dinucleotide distribution with a first-order Markov model on the test species *S. cerevisiae*. 3' UTRs are known to exhibit strong dinucleotide biases, making this model a realistic and hard-to-beat baseline.

As a first test, we analyzed to what extent our model could reconstruct masked nucleotides in the annotated 3' UTR sequences of the test species *Saccharomyces*

cerevisiae. In random regions of these 3' UTR sequences the reconstruction fidelity, as measured by the masked nucleotide probability, was similar to what is achieved using the dinucleotide distribution (Fig 2B). Perhaps, this reflects our current understanding that 3 'UTRs are not densely packed with RNA motifs but consist of few motifs spaced by sequences merely characterized by their dinucleotide frequencies. Notably, the species-agnostic 11-mer model did poorer than the species-specific dinucleotide model on random regions, reflecting the large variation of dinucleotide preferences in 3' UTR across the evolutionary tree.

Importantly, reconstruction improved significantly on regulatory sites (Fig. 2B). Specifically, we considered consensus motifs of six RNA binding proteins, as well as binding sites identified by a gPAR-CLIP experiment. Our species-aware language model reconstructed instances of all these regulatory sequences better than the dinucleotide or species-agnostic model, except for the Nrd1 motif, where it was on-par with the species-agnostic model. Nevertheless, on the highly conserved motif Puf3, even the species-agnostic model reconstructed significantly better than the dinucleotide model. These results show that the masked nucleotide modeling approach prioritizes regulatory elements. Moreover, this shows that encoding species information enables improved generalization to regulatory motifs in unseen species.

Compared to DNABERT, which has 100x more parameters, our species-aware model performed remarkably well. It reconstructed in median better than DNABERT for 3 out of 6 motifs and for gPAR-CLIP bound sites. For the other 3 motifs, there was no statistically significant difference in performance. This clearly shows the benefits of species-awareness. Especially, as DNABERT does generally outperform our species-agnostic model, likely as a result of the much larger parameter count.

Species-aware model captures Puf3 motif evolution

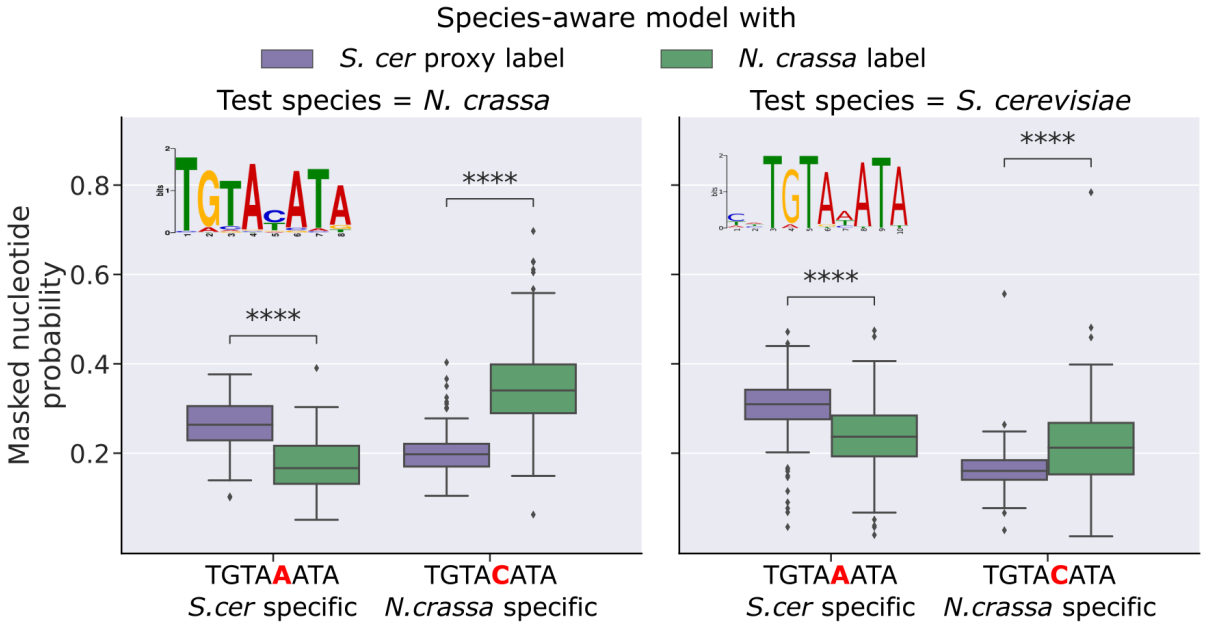

Fig. 3: Species-awareness recovers Puf3 motif evolution. Reconstruction performance (masked nucleotide probability) of the fifth nucleotide of Puf3 motif instances in *N. crassa* and *S. cerevisiae* 3' regions.

Next, we asked whether our model could allow studying the evolution of regulatory motifs. To this end, we considered the specific case of the Puf3 motif.

Puf3 homologues are found in virtually all eukaryotes and its binding motif is highly conserved (Hogan *et al.*, 2015). However, the consensus binding sequence of Puf3 has slightly evolved between *S. cerevisiae*, preferring TGTA**A**ATA, and the filamentous fungus *Neurospora crassa* preferring TGTA**C**ATA (Wilinski *et al.*, 2017). Remarkably, our model captured this evolution as it predicts a C at the +5 position when told to predict for *N. crassa* and an A when told to predict for our *S. cerevisiae* proxy *C. glabrata* (Fig. 3). This example shows that our species-aware model, the first of its kind, has learned to some extent to account for the evolution of the regulatory code.

Species-aware masked nucleotide modeling enables direct prediction of experimentally validated binding sites

Method	Binding vs. Nonbinding	Binding vs. Shuffled	PUF3 Binding vs. Nonbinding
Conservation	0.530 (0.013)	-	**0.682 (0.074)**
DNABERT	0.646 (0.012)	0.564 (0.009)	0.562 (0.08)
Species-agnostic	0.672 (0.012)	0.581 (0.009)	0.533 (0.08)
Species-aware	**0.689 (0.012)** *	**0.589 (0.009)** *	0.620 (0.077)

Table 1: Masked nucleotide probability predicts protein-bound nucleotides. AUC and 2 standard deviations thereof. Best model in bold, star denotes statistical significance compared to the next best model ($P < 0.05$, DeLong test).

To further assess how well masked nucleotide probability reflects function, we considered regions reported to be protein-bound as identified by gPAR-CLIP, as a proxy for regions of regulatory importance. Using the masked nucleotide probability as a predictor distinguished bound from unbound nucleotides with AUC 0.69 (Table 1), which is significantly better than chance and also outperforms the phastCons score, a conservation score derived from whole-genome alignment of six yeast species to *S. cerevisiae* (Siepel *et al.*, 2005).

Different reconstruction for binding and nonbinding regions may be an indirect consequence of large differences in nucleotide composition. To control for this, we reframe the task as distinguishing bound regions from di-nucleotide shuffled versions thereof. We again find that the model can discriminate better than chance, which demonstrates that the classification performance is not simply a consequence of GC or dinucleotide content. Notably, some di-nucleotide shuffled bound regions may themselves be genuinely bound, thus the classification performance reported may be an underestimate.

As an additional task, we asked whether the model could distinguish between Puf3 consensus motif sites which occur in the 3' UTR of genes which have been experimentally confirmed to be bound by Puf3 from those which occur in genes which show no experimental evidence of Puf3 binding (Lapointe *et al.*, 2018). Again, we find that our model can do so better than chance. Moreover, while the conservation score performs better, this difference is not significant. Remarkably the conservation score was made using whole genome alignments of Saccharomyces species, thus it used genomes significantly more closely related to *S. cerevisiae* than used by our approach.

Admittedly, the predictive power of our model on these tasks is overall mediocre. However, it was never trained to predict protein binding, has never even seen this species, and was not further tuned for this prediction task. Therefore, the preceding

analysis demonstrates that our approach does capture some notion of the regulatory importance of a sequence without requiring any experimental labels during training.

Embeddings enable state-of-the-art mRNA stability predictions

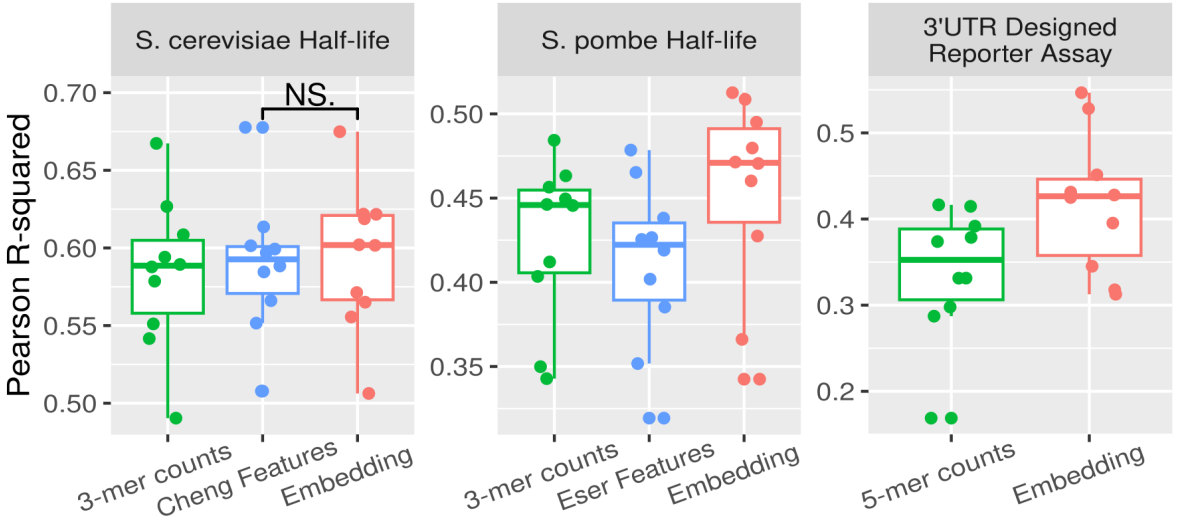

Fig. 4: 3' UTR embeddings of the species-aware model outperform other methods on a variety of downstream tasks. In each case, performance for 10 cross-validation folds is shown. Embeddings from the species-aware model are compared with state-of-the-art models, namely (Cheng *et al.*, 2017) and (Eser *et al.*, 2016), and with k-mer count regressions, where the best k from {1,...,7} is shown. The embedding model significantly outperforms ($P < 0.05$) all other approaches, except Cheng on the S. cerevisiae mRNA half-life task.

The ability of the model to predict regulatory regions directly from the masked nucleotide probability without any further training is remarkable, but the predictions do not yet reach the quality required for practical use. An established way to use masked language models for downstream supervised prediction is to extract the so-called embedding, i.e. the high-dimensional representations upon which the masked predictions are made. For our models, these are single-nucleotide level representations. By pooling these nucleotide-level representations along a sequence, we can embed entire 3' UTR sequences, thus creating a numerical summary of the contained nucleotides.

We computed such embeddings for *S. cerevisiae* 3' UTR sequences and used them, in addition to some other features derived from the CDS and 5'UTR, in a simple linear model to predict the measured mRNA half-life (Sun *et al.*, 2013). We found that using our embeddings as representation of the 3' UTR outperformed using k-mer-counts (bag of words approach) and performed on par with the state-of-the-art model (Cheng *et al.*, 2017). Note that this model encodes the 3' UTR as motif counts which were found using a complicated multi-stage selection procedure. In contrast, our embeddings require essentially no finetuning or feature engineering.

We performed a similar analysis for *S. pombe.* Again we found that using our embeddings as representation of the 3' UTR outperforms k-mer based approaches. Moreover, in the case of *S. pombe*, we also significantly outperformed the state-of-the-art model (Eser *et al.*, 2016).

Lastly, we considered a designed reporter experiment, where the protein expression of different mRNA molecules was measured (Shalem *et al.*, 2015). In this

experiment, the 5'UTR and CDS of molecules was constant and only the 3' UTR varied. Once more our approach significantly outperformed other methods of encoding the 3' UTR sequence.

Overall, this analysis shows that embeddings of 3' UTR sequences provided by our model can be used by simple linear downstream models to effectively predict molecular phenotypes, without requiring complicated finetuning or feature engineering. To our knowledge, this is the first demonstration of sequence embeddings capable of matching or outperforming state-of-the-art supervised models for predicting a RNA metabolic rate.

Discussion

Here we showed that a species-aware masked nucleotide language model trained on a large corpus of species genomes can be used to reconstruct known RNA binding protein consensus motifs significantly better than chance and species-agnostic models. Moreover, the model was able to distinguish experimentally validated binding regions in *S. cerevisiae* 3' untranslated regions from background, despite never having been trained for this task and never having seen a single member of the Saccharomyces genus during training. Additionally, embeddings derived from the model could be used in a linear regression to achieve prediction of the impact of the 3'UTR on mRNA half-life. This way we were able to perform on-par with the state-of-the-art model in *S. cerevisiae*, which employed extensive manual feature selection, and established a new state-of-the-art for *S. pombe*. Overall, our results demonstrate that the masked nucleotide modeling approach has allowed capturing and meaningfully embedding aspects of the regulatory code of simple eukaryotes. We foresee that the same approach could be applied to 5' UTR and promoter regions and to higher eukaryotes or other domains of life.

To our knowledge, we are the first to investigate the ability of a masked language model to reconstruct noncoding DNA (specifically gene 3' prime regions) in a fully held-out genus. While our models do show success at this task, it is important to note that the reconstruction fidelity achieved is far lower than what has been reported for natural language (Devlin *et al.*, 2019). Letting aside the massive amount of training data available in NLP, one reason for this difference is that selective constraints on individual nucleotides in non-coding regions are low because non-coding single nucleotide variants often do not affect regulation, whereas changes of a single word in a natural language sentence often strongly affect meaning when it does not offend syntax or grammar. In other words, the masked nucleotide prediction task is inherently more ambiguous than the masked token modeling in NLP. There may be a tight upper bound on the overall nucleotide reconstruction fidelity which can be achieved and high-fidelity reconstruction of *unconstrained* DNA is not particularly desirable in any case. Nevertheless, we expect further progress on reconstructing elements with regulatory importance.

We are also the first to make a DNA language model species-aware. We show that this significantly improves the ability of the model to generalize. Moreover, it opens the possibility of using the model to study the evolution of regulatory elements, by analyzing how predictions change as we change the species label. We have made a first foray in this direction by showing that the model captures a change in the Puf3 binding motif between two highly diverged fungi. We expect that future

work will use similar approaches in more sophisticated language models to further trace the evolution of the regulatory code.

Parallel to our work, another study was released which also trains on genomes of multiple species and focuses on human for downstream tasks (Dalla-Torre *et al.*, 2023). The proposed model is species-agnostic but more than one thousand times larger than ours. Unfortunately, as of writing, the model was not made publically available, therefore we cannot currently judge whether such huge parameter counts suffice to overcome the drawbacks of lacking species information.

Lastly, we present the first use of a state space model for language modeling in genomics. In comparison to transformers, the de-facto state-of-the-art for language modeling in natural language, state space models have been shown to scale very well to sequences of arbitrary length while nevertheless capturing long-range dependencies (Gu *et al.*, 2021). In principle, this should make them perfectly suited for non-coding DNA sequences, especially in higher eukaryotes where elements can interact that are separated by hundreds of thousands of bases.

Acknowledgements

This study was funded by the German Bundesministerium für Bildung und Forschung (BMBF) through the Model Exchange for Regulatory Genomics project MERGE (031L0174A to JG). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

We thank all the scientists who isolated, sequenced and assembled the genome data which builds the basis for our research.

Bibliography

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Benegas,G. *et al.* (2022) DNA language models are powerful zero-shot predictors of non-coding variant effects. 2022.08.22.504706.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cheng,J. *et al.* (2017) Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA*, **23**, 1648–1659.

Dalla-Torre,H. *et al.* (2023) The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. 2023.01.11.523679.

Devlin,J. *et al.* (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Eraslan,G. *et al.* (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, **20**, 389–403.

Eser,P. *et al.* (2016) Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol. Syst. Biol.*, **12**, 857.

Foat,B.C. *et al.* (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci.*, **102**, 17675–17680.

Freeberg,M.A. *et al.* (2013) Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. *Genome Biol.*, **14**, R13.

Gu,A. *et al.* (2021) Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers.

Gupta,A. *et al.* (2022) Diagonal State Spaces are as Effective as Structured State Spaces.
Hogan,D.J. *et al.* (2008) Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLOS Biol.*, **6**, e255.
Hogan,G.J. *et al.* (2015) Evolutionary Conservation and Diversification of Puf RNA Binding Proteins and Their mRNA Targets. *PLOS Biol.*, **13**, e1002307.
Ji,Y. *et al.* (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112–2120.
Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
Kimura,M. (1968) Evolutionary Rate at the Molecular Level. *Nature*, **217**, 624–626.
Lapointe,C.P. *et al.* (2018) Multi-omics Reveal Specific Targets of the RNA-Binding Protein Puf3p and Its Orchestration of Mitochondrial Biogenesis. *Cell Syst.*, **6**, 125-135.e6.
Pelechano,V. *et al.* (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
Prieto,M. and Wedin,M. (2013) Dating the Diversification of the Major Lineages of Ascomycota (Fungi). *PLOS ONE*, **8**, e65576.
Shalem,O. *et al.* (2015) Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. *PLOS Genet.*, **11**, e1005147.
Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
Sun,M. *et al.* (2013) Global Analysis of Eukaryotic mRNA Degradation Reveals Xrn1-Dependent Buffering of Transcript Levels. *Mol. Cell*, **52**, 52–62.
Wilinski,D. *et al.* (2017) Recurrent rewiring and emergence of RNA regulatory networks. *Proc. Natl. Acad. Sci.*, **114**, E2816–E2825.
Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.