Ab Initio Functional Annotation of a Novel DNA Sequence from a Newly Discovered Species

GROUP MEMBERS
ABIN MATHEW [MS21087]
FAYIZ M [MS21078]
NEHA KOONERI [MS21207]
SHUBHAM LADHWAL[MS21205]

November 19, 2024

Table of Contents

List of Figures

1 INTRODUCTION

1.1 Ab-Initio Method

The emergence of next-generation sequencing (NGS) technologies transformed genomics and molecular biology. This leads to a rapid and precise sequencing of DNA from a diverse range of organisms. However, some problems exist in functionally annotating novel DNA sequences, particularly for species or strains that are not represented in existing genomic databases. The absence of homologous sequences, as in the newly discovered species, creates gaps in understanding gene functions and biological roles. Accurate functional annotation is critical in uncoiling the genetic basis of traits, processes, and interactions. Traditional annotation methods rely on homology-based techniques, which depend on existing genomic databases. However, in the case of highly divergent species, these approaches might fail and lead to incomplete annotations. We should look for computational tools to annotate novel sequences in such a situation.[1]

The word Ab initio means "From scratch". The ab initio method focuses on intrinsic sequence features rather than known homologs. It involves the identification of gene characteristics such as open reading frames (ORFs), codon usage patterns, and splice site motifs. For species lacking a genetic database, the Ab initio method can be used to predict the coding regions, gene structure, and regulatory elements. The Ab Initio method for functional annotation is a computational approach used in genomics to predict gene function without relying on existing experimental data or annotation databases. Despite their potential, ab initio prediction tools vary in accuracy and reliability. [2]

1.2 Objective

The problem is
You found a new species and extracted the DNA. After sequencing, you found a novel previously unknown chunk of sequence (unknown_sequence.fna). **Functionally annotate this sequence using ab initio methods and NOT through homology-based methods.**

Investigating the genetic foundations of novel organisms can provide profound insights into evolutionary mechanisms, support the discovery of new biomolecules for medical and biotechnological applications, and deepen our understanding of biodiversity. Our project is focused on addressing a critical question: How can we effectively utilize ab initio methods to functionally annotate a previously uncharacterized DNA sequence from a novel organism? By tackling this problem, we aim to enhance our understanding of gene function in unexplored species and contribute to advancing knowledge in the field of genomics, ultimately bridging gaps in our molecular understanding of biodiversity.

2 METHODS

In order to tackle the question, we used several techniques that were taught in class like T code analysis and motif score assessment. We also used Augustus.

Given Sequence: Unknown sequence2.fna ($> seq2$)

A snapshot of the given unknown sequence file is given below.

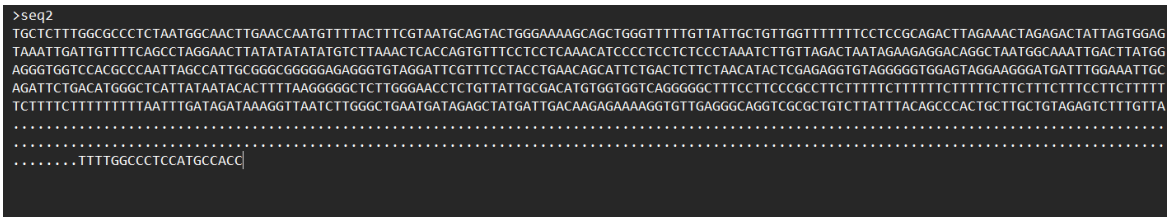

Figure 1: Unknown sequence

2.1 T Code Analysis and CpG ratio

Fickett's T-Code, developed by James Fickett in 1982, is an algorithm designed to identify protein-coding regions in DNA sequences. It determines whether a DNA fragment is coding or non-coding by analyzing nucleotide composition, positional preference, and sequence periodicity. Coding sequences display unique patterns in their adenine, thymine, cytosine, and guanine distributions, and exhibit a period-3 signal reflecting their triplet codon structure. The algorithm involves calculating frequency ratios and positional biases of nucleotides to detect three-base periodicity. Based on these analyses, the T-Code assigns a score indicating the likelihood of a segment being coding, with a higher score suggesting a coding segment and a lower score indicating a non-coding one.[3]

The positional bias for A can be derived by,

$$P_A = \frac{\max(A_1, A_2, A_3)}{\min(A_1, A_2, A_3) + 1} \tag{1}$$

$$T = \sum_{i=1} \omega_i \cdot X_i \tag{2}$$

Here ω is the weight values. X is the compositional and positional parameters

The CpG ratio, comparing observed to expected CpG dinucleotide frequencies, is key in identifying CpG islands (CGIs). Traditional methods, like Gardiner-Garden criteria, use thresholds for GC content ($\geq 50\%$), CpG ratio (≥ 0.6), and length ($\geq 200bp$) but may miss short functional regions. High CpG ratios are characteristic of CGIs, which are often found in gene promoters and regulatory regions

Tools like CpGcluster improve CGI detection by analyzing CpG density statistically, enabling identification of smaller regions crucial for gene regulation and epigenetic studies, especially in cancer research.traditional ab initio methods heavily rely on the CpG ratio, modern algorithms like CpGcluster expand CGI detection by focusing on statistical properties, providing a more comprehensive approach to uncovering CpG-rich regions and their functional roles.[4][5]

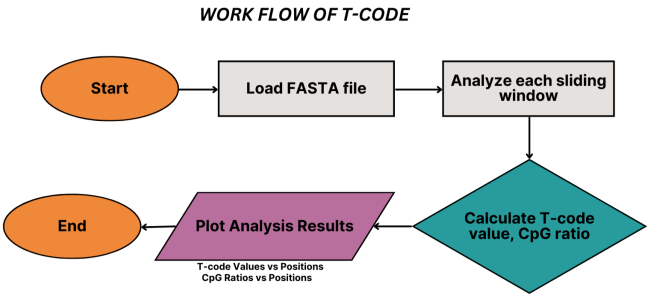

Figure 2: T code workflow

2.2 Augustus

Gene prediction tools have improved significantly, but their effectiveness is still limited in many genomic projects. A key method to enhance gene structure predictions is integrating outputs from various gene finders with data from expressed sequence tags (ESTs) and protein sequences. AUGUSTUS, using a Generalized Hidden Markov Model (GHMM), excels in this realm by incorporating extrinsic data, which boosts prediction accuracy. Recent updates enable AUGUSTUS to predict multiple transcripts per gene, addressing the complexities of alternative splicing in human genes while allowing researchers to fine-tune sensitivity and specificity. The tool segments DNA sequences into exons, introns, and intergenic regions, and employs random sampling to generate alternative splice variants. Its user-friendly web interface accepts FASTA format sequences and provides outputs that detail exon and intron boundaries and predicted protein sequences. Performance evaluations, including those from the ENCODE project, indicate that predicting more transcripts enhances gene sensitivity while maintaining specificity.AUGUSTUS stands out compared to other gene prediction tools, particularly in identifying multiple splice variants, thus aiding gene annotation efforts in genomic research.[6]

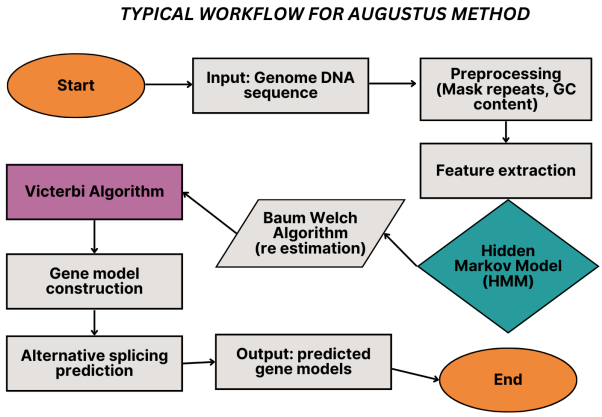

Figure 3: AUGUSTUS workflow

2.3 Motif score

JASPAR-style Position-Specific Scoring Matrices (PSSMs) are utilized to evaluate motifs within DNA sequences. A PSSM is a matrix that indicates nucleotide frequencies or log-likelihoods at each motif position. Scoring involves aligning each nucleotide in a sequence with the corresponding motif position, summing the scores to assess similarity between the sequence and the motif. Although typically not part of the Ab Initio method, it has been used here as a confirmation step. JASPAR is a repository of transcription factor binding profiles, offering a collection of position frequency matrices (PFMs) that are utilized to investigate gene regulation and the interactions between transcription factors and DNA.

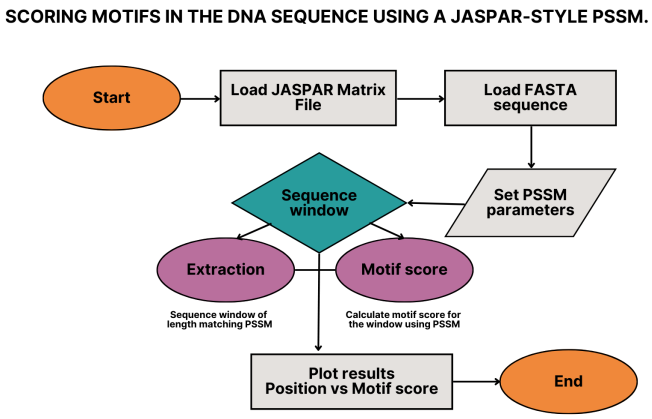

Figure 4: T code workflow

2.4 Programming and GitHub file

We used Python programming to execute the T code, CpG, and motif score analysis. Here is a link to the GitHub repository that contains Python scripts, sequence files, Results and other relevant materials for the project.

GitHub Repository

3 RESULTS

3.1 Results from T code analysis and CpG ratio

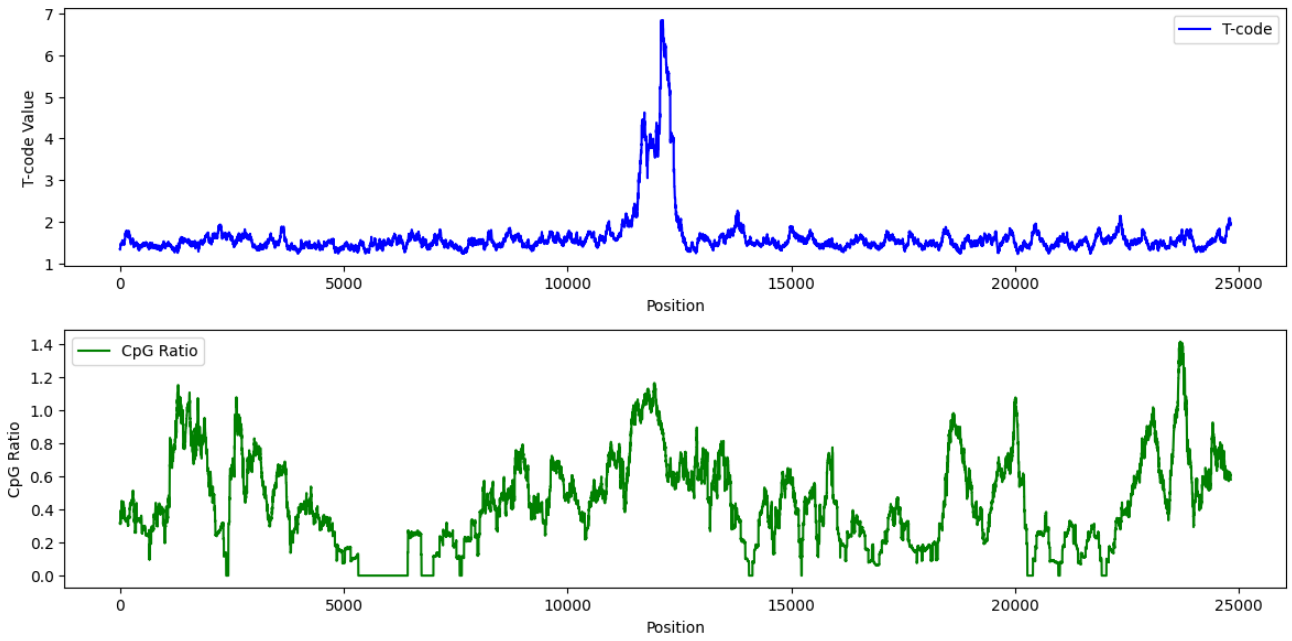

Figure 5: T code and CpG

Position with maximum T-code: 12131
Maximum T-code value: 6.8414
Corresponding CpG ratio: 0.8619

- A visible peak can be observed at position 12131, this indicates regions of higher coding potential. This might suggest possible protein-coding regions, exons, or transcriptionally active regions.
- Area with high CpG ratio might indicate regulatory activity, promoter regions for genes, transcriptional initiation and targets of epigenetic modifications
- Overall the T-code and CpG ratio analysis suggest that our unknown sequence contains potential coding regions and regulatory elements.

3.2 Results from Motif scores

We looked at three different JASPAR frequency matrices (ID MA0108.1,ID MA0108.2,ID MA0108.3) to assign motif scores throughout the sequence. These matrices were not species-specific. The results are given below;

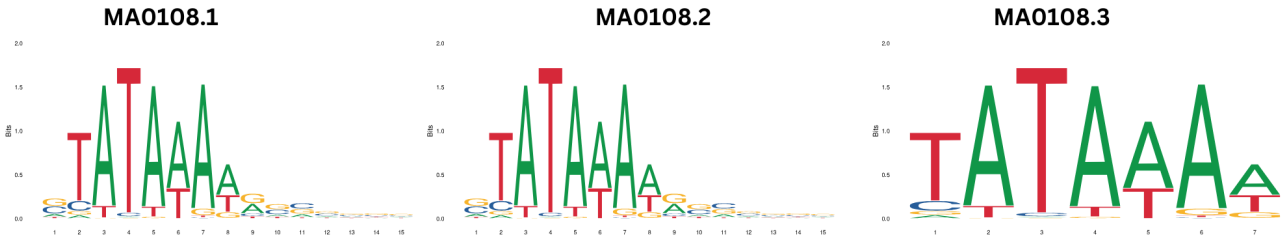

Figure 6: Frequency matrix

MATRIX ID	POSITION WITH MAXIMUM SCORE	MAXIMUM SCORE
MA0108.1	12060	9.41×10^{-24}
MA0108.2	12060	9.17×10^{-24}
MA0108.3	6460	2.78×10^{-7}

Table 1: JASPAR Frequency matrix

From all these results, we can verify that the matrix ID MA0108.1 is aligned with the results of the CpG and T code. Based on matrix ID MA0108.1 we went for matrix score across sequence. The graph is plotted below.[7]

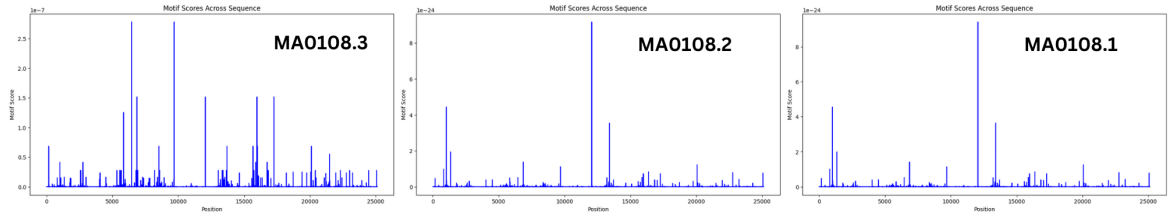

Figure 7: Motif score

- There is a significant peak at a specific position at 12060, indicating a highly scored motif. This could correspond to a strong transcription factor binding site or a regulatory element.
- If linked with high T-code values and CpG sites, this region is likely to be involved in coding sequences or might play a role in transcriptional regulation by binding certain transcription factors.

By integrating results from the T code, the CpG analysis, and motif scoring, it appears that the unknown DNA sequence likely contains important functional elements, including protein-coding regions and regulatory motifs.

3.3 Results from Augustus

Using AUGUSTUS (Version 3.3.3) the following results were obtained from the analysis of the given unknown DNA sequence. The analysis was performed using a human species model with UTR detection disabled. [8]

```
# ----- prediction on sequence number 1 (length = 25120, name = seq2) -----
#
# Predicted genes for sequence number 1 on forward strand
# start gene g1
seq2    AUGUSTUS        gene    11741   12694   0.98    +       .       g1
seq2    AUGUSTUS        transcript      11741   12694   0.98    +       .       g1.t1
seq2    AUGUSTUS        start_codon     11741   11743   .       +       0       transcript_id "g1.t1"; gene_id "g1";
seq2    AUGUSTUS        single  11741   12694   0.98    +       0       transcript_id "g1.t1"; gene_id "g1";
seq2    AUGUSTUS        CDS     11741   12694   0.98    +       0       transcript_id "g1.t1"; gene_id "g1";
seq2    AUGUSTUS        stop_codon      12692   12694   .       +       0       transcript_id "g1.t1"; gene_id "g1";
# coding sequence = [atgtacaacatgatggagacggagctgaagccgccgggcccgcagcaaacttcgggggggcggcggcggcaactccaccg
# cggcggcggccggcggcaaccagaaaaacagcccggaccgcgtcaagcggcccatgaatgccttcatggtgtggtcccgcgggcagcggcgcaagatg
# gcccaggagaaccccaagatgcacaactcggagatcagcaagcgcctgggcgccgagtggaaacttttgtcggagacggagaagcggccgttcatcga
# cgaggctaagcggctgcgagcgctgcacatgaaggagcacccggattataaataccggccccggcggaaaaccaagacgctcatgaagaaggataagt
# acacgctgcccggcgggctgctggcccccggcggcaatagcatggcgagcggggtcggggtgggcgccggcctgggcgcgggcgtgaaccagcgcatg
# gacagttacgcgcacatgaacggctggagcaacggcagctacagcatgatgcaggaccagctgggctacccgcagcacccgggcctcaatgcgcacgg
# cgcagcgcagatgcagcccatgcaccgctacgacgtgagcgccctgcagtacaactccatgaccagctcgcagacctacatgaacggctcgcccacct
# acagcatgtcctactcgcagcagggcacccctggcatggctcttggctccatgggttcggtggtcaagtccgaggccagctccagccccctgtggtt
# acctcttcctcccactccagggcgccctgccaggccggggacctccgggacatgatcagcatgtatctccccggcgccgaggtgccggaacccgccgc
# ccccagcagacttcacatgtcccagcactaccagagcggcccggtgcccggcacggccattaacggcacactgccctctcacacatgtga]
# protein sequence = [MYNMMETELKPPGPQQTSGGGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAE
# WKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD
# QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMALGSMGSVVKSEASSSPPVVTSSSHSRAPCQAGDLRDMI
# SMYLPGAEVPEPAAPSRLHMSQHYQSGPVPGTAINGTLPLSHM]
```

Figure 8: Results from Augustus

- Predicted gene: Gene $g1$ (Sequence: $seq2$, forward strand)
- Location: Base pairs 11,741 to 12,694.
- Confidence score: 0.98.

Gene structure

- Transcript ID: $g1.t1$
- Start Codon: Base pairs 11,741 to 11,743
- Stop Codon: Base pairs 12,692 to 12,694
- Coding Sequence (CDS): Base pairs 11,741 to 12,694

Observations

- Protein Length: 318 amino acids
- The prediction identifies a single gene ($g1$) on the forward strand with high confidence
- Coding sequence length: 954

4 DISCUSSION

The advent of next-generation sequencing (NGS) has revolutionized genomics, enabling rapid and high-throughput sequencing while offering insights into genetic diversity. In this project, we used ab initio methods to functionally annotate a novel DNA sequence, to provide predictions of gene characteristics, such as open reading frames (ORFs) and regulatory motifs, even in the absence of prior sequence knowledge or databases. This capability is particularly beneficial for species that lack substantial genomic databases, allowing a deeper exploration of their genetic architecture.

Assessment of Coding Potential: We utilized Fickett's T-Code to evaluate the coding potential of the unknown DNA sequence. This method identified regions with elevated T-Code values (*Position with maximum T-code: 12131,Maximum T-code value: 6.8414*). These peaks align with the expected nucleotide composition and the characteristic patterns of protein-coding sequences, providing strong indicators of functional genomic elements in the sequence.

Regulatory Insights from CpG Ratio Analysis: The CpG ratio was analyzed to understand regulatory activities, often associated with gene promoter regions. Elevated CpG (*CpG ratio: 0.8619*) ratios were correlated with areas of significant coding potential, suggesting that these regions may have regulatory significance or potential information.

Transcription Factor Binding Site Prediction: By applying JASPAR-style position-specific scoring matrices (PSSMs), we identified probable transcription factor binding sites (*significant peak at position: 12060*). This analysis offers valuable information on the regulatory control mechanisms underlying gene expression within the novel sequence.

Integrated Analysis for Enhanced Annotation: Integrating T-Code peaks, CpG ratio data, and transcription factor binding motif scores allowed us to uncoil the coding features, protein coding and transcriptional regulation of the sequence. These results suggest a critical role for these regions in transcriptional regulation and gene expression within the newly characterized species.

Gene Prediction and Structural Analysis: The use of Augustus, a gene prediction tool, further supported our annotation efforts. Augustus predicted coding sequences (*Gene $g1$ between 1,741 to 12,694 base pairs*) and identified transcripts and gene structure. The high-confidence prediction (*confidence score: 0.98*) of gene $g1$ underscores its reliability in delineating gene structure, making it a vital component of our analysis pipeline.

Limitations and Future Directions: Despite the strength of these computational predictions, experimental validation remains essential to confirm gene function and regulatory roles. Ab initio methods, while powerful, are inherently limited by their reliance on computational algorithms without homologous sequence references. Future efforts will focus on validating these predictions through experimental techniques such as RT-PCR, ChIP-Seq, or RNA-Seq to confirm transcriptional activity and regulatory interactions.

5 CONCLUSION

Our Project explores the critical role of ab initio methods for functional annotation in genomics, particularly when addressing uncharacterized sequences from novel species. By integrating T-Code analysis, CpG ratio assessments, motif scoring, and advanced gene prediction algorithms like Augustus, we made a good approach to uncover the functional elements within the unknown sequence. This multifaceted methodology not only enhances our understanding of gene function in unexplored organisms but also contributes significantly to the fields of biodiversity and evolutionary biology.

As we continue to investigate the genetic foundations of newly discovered species, this Project provides some groundwork for future studies aimed at identifying novel bio molecules with applications in medicine and biotechnology. The urgent need for accurate functional annotations in genomics remains essential to bridging gaps in our molecular understanding of life's complexities and processes. Overall, our work validates the effective use of ab initio approaches and emphasizes their importance in advancing our idea of genetic frameworks, ultimately shaping the future area of genomics research and its potential for groundbreaking discoveries.

5.1 Rubric table of relative weights

Member	Understanding	Analysis	Ideation/discussion	Resourcefulness	Writing/presentation	Co-operation
MS21087	0.25	0.25	0.25	0.25	0.25	0.25
MS21078	0.25	0.25	0.25	0.25	0.25	0.25
MS21207	0.25	0.25	0.25	0.25	0.25	0.25
MS21205	0.25	0.25	0.25	0.25	0.25	0.25

Table 2: Rubric sheet

References

[1] G. F. Ejigu and J. Jung, "Review on the computational genome annotation of sequences obtained by next-generation sequencing," *Biology*, vol. 9, no. 9, 2020.

[2] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, pp. e132–e132, 04 2010.

[3] J. W. Fickett, "Recognition of protein coding regions in dna sequences," *Nucleic Acids Research*, vol. 10, pp. 5303–5318, 09 1982.

[4] B. G. C. P. L.-E. P. L. P. C. O. J. L. Hackenberg, Michael, "Cpg islands or cpg clusters: how to identify functional gc-rich regions in a genome?," *BMC Genomics*, vol. 11, 05 2010.

[5] Z. Z. Han, Leng, "Cpg islands or cpg clusters: how to identify functional gc-rich regions in a genome?," *BMC Bioinformatics*, vol. 10, 02 2009.

[6] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "Augustus: ab initio prediction of alternative transcripts," *Nucleic Acids Research*, vol. 34, pp. W435–W439, 07 2006.

[7] I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu, J. Castro-Mondragon, K. Ferenc, V. Kumar, R. B. Lemma, J. Lucas, J. ChÃ¨neby, D. Baranasic, A. Khan, O. Fornes, S. Gundersen, M. Johansen, E. Hovig, B. Lenhard, A. Sandelin, W. Wasserman, F. Parcy, and A. Mathelier, "Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 52, pp. D174–D182, 11 2023.

[8] "Web augustus service: University of greifswald," 2019.