UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Machine Learning Algorithm for Predicting Translation Initiation Start Positions

Diploma Thesis

Digenis Stefanos

Supervisor: Dimitrios Katsaros

October 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Machine Learning Algorithm for Predicting Translation Initiation Start Positions

Diploma Thesis

Digenis Stefanos

Supervisor: Dimitrios Katsaros

October 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αλγόριθμος Μηχανικής Μάθησης για την Πρόβλεψη Θέσεων Έναρξης της Μετάφρασης

Διπλωματική Εργασία

Διγενής Στέφανος

Επιβλέπων: Δημήτριος Κατσαρός

Οκτώβριος 2023

Approved by the Examination Committee:

Supervisor **Dimitrios Katsaros**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Artemis Xatzigeorgiou**
Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

Member **Dimitrios Rafailidis**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Mrs. Xatzigeorgiou Artemis, for entrusting me with this incredibly engaging and captivating research topic. Her guidance, support, and invaluable insights have been instrumental in the successful completion of this thesis.

I would also like to extend my heartfelt appreciation to Mrs. Xatzigeorgiou Artemis's previous Ph.D. student, Dimitris Grigoriadis, for his generous assistance throughout this journey. His mentorship, encouragement, and sage advice have played a pivotal role in shaping the direction and quality of my work.

Furthermore, I am indebted to my esteemed professors, Dimitris Katsaros and Dimitris Rafailidis, for their gracious acceptance of my thesis topic and their willingness to serve on my examination committee. Their expertise and constructive feedback have enriched my research and contributed to its academic rigor.

I am deeply grateful to all those who have supported me in this academic endeavor, and I look forward to continuing my scholarly pursuits with the knowledge and experience gained during this thesis.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Digenis Stefanos

Diploma Thesis

Machine Learning Algorithm for Predicting Translation Initiation Start Positions

Digenis Stefanos

Abstract

This thesis focuses on the development of a state-of-the-art deep learning model utilizing convolutional neural networks (CNNs) for the classification of RNA sequences. Specifically, the model aims to predict the presence or absence of translation initiation start sites within these sequences. Translation initiation plays a crucial role in protein synthesis, making accurate prediction of start sites a significant challenge in computational biology.

The proposed deep learning model leverages the power of CNNs to effectively capture relevant features in the RNA sequences. By training the model on a comprehensive dataset of annotated translation initiation sites, it learns to discern key patterns and motifs associated with initiation starts. Through an extensive evaluation process, the performance of the model has demonstrated a remarkably high accuracy rate, surpassing existing tools in this domain.

The significance of this research lies in the potential impact on various biological applications, including gene annotation, drug target identification, and understanding the mechanisms of protein synthesis. The ability to reliably predict translation initiation start sites using a deep learning approach provides a valuable tool for both computational biologists and experimental researchers.

Keywords:

deep learning, convolutional neural networks, RNA sequences, translation initiation start prediction, computational biology.

Διπλωματική Εργασία

Αλγόριθμος Μηχανικής Μάθησης για την Πρόβλεψη Θέσεων Έναρξης της Μετάφρασης

Διγενής Στέφανος

Περίληψη

Η παρούσα διατριβή επικεντρώνεται στην ανάπτυξη ενός μοντέλου βαθιάς μάθησης τελευταίας τεχνολογίας που χρησιμοποιεί συνελικτικά νευρωνικά δίκτυα (CNN) για την ταξινόμηση αλληλουχιών RNA. Συγκεκριμένα, το μοντέλο αποσκοπεί στην πρόβλεψη της παρουσίας ή της απουσίας θέσεων έναρξης της μετάφρασης εντός αυτών των αλληλουχιών. Η έναρξη της μετάφρασης διαδραματίζει κρίσιμο ρόλο στην πρωτεϊνοσύνθεση, καθιστώντας την ακριβή πρόβλεψη των θέσεων έναρξης μια σημαντική πρόκληση στην υπολογιστική βιολογία.

Το προτεινόμενο μοντέλο βαθιάς μάθησης αξιοποιεί τη δύναμη των CNN για την αποτελεσματική σύλληψη των σχετικών χαρακτηριστικών στις ακολουθίες RNA. Εκπαιδεύοντας το μοντέλο σε ένα ολοκληρωμένο σύνολο δεδομένων με σχολιασμένες θέσεις έναρξης μετάφρασης, μαθαίνει να διακρίνει βασικά μοτίβα που σχετίζονται με τις θέσεις έναρξης μετάφρασης. Μέσω μιας εκτεταμένης διαδικασίας αξιολόγησης, οι επιδόσεις του μοντέλου κατέδειξαν ένα αξιοσημείωτα υψηλό ποσοστό ακρίβειας, ξεπερνώντας τα υπάρχοντα εργαλεία στον τομέα αυτό.

Η σημασία αυτής της έρευνας έγκειται στον πιθανό αντίκτυπο σε διάφορες βιολογικές εφαρμογές, όπως ο σχολιασμός γονιδίων, ο εντοπισμός στόχων φαρμάκων και η κατανόηση των μηχανισμών της πρωτεϊνοσύνθεσης. Η ικανότητα αξιόπιστης πρόβλεψης των θέσεων έναρξης της μετάφρασης με τη χρήση μιας προσέγγισης βαθιάς μάθησης παρέχει ένα πολύτιμο εργαλείο τόσο για τους υπολογιστικούς βιολόγους όσο και για τους πειραματικούς ερευνητές.

Λέξεις-κλειδιά:

βαθιά μάθηση, συνελικτικά νευρωνικά δίκτυα, αλληλουχίες RNA, πρόβλεψη έναρξης μετάφρασης, υπολογιστική βιολογία.

Table of contents

List of figures

List of tables

Abbreviations

TIS Translation Initiation Start

ORF Open Reading Frames

uORF Upstream Open Reading Frames

RNA Ribonucleic acid

mRNA Messenger Ribonucleic acid

tRNA Transfer Ribonucleic acid ribosomal

rRNA Ribosomal Ribonucleic acid

DNA Deoxyribonucleic acid

IRES Internal Ribosome Entry Sites

GFT General Transfer Format

FTP File Transfer Protocol

BED Browser Extensible Data

bps Base Pairs

Kbps Kilo Base Pairs

IGV Integrative Genomics Viewer

UCSC University of California Santa Cruz

ML Machine Learning

SVM Support Vector Machine

ANN Artificial Neural Network

RNN Recurrent Neural Network

CNN Convolutional Neural Network

MLP Multi-layer Perceptron

relu Rectified Linear Unit

selu Scaled Exponential Linear Unit

TP True Positives

TN True Negatives

FP False Positives

FN False Negatives

Chapter 1

Introduction

In the realm of molecular biology and genomics, the prediction of translation initiation start positions in RNA sequences plays a fundamental role in unraveling the mysteries of gene expression and protein synthesis. This thesis aims to address the challenges associated with accurately identifying these start codons by developing a machine-learning algorithm specifically designed for this task. Before delving into the specific details of the algorithm and its architecture, it is essential to provide a general description of the scope of this thesis.

The focus of this research lies within the realm of genomic analysis, where the intricate nature of RNA sequences presents both opportunities and challenges. RNA sequences, composed of nucleotide bases, hold the key to understanding how genetic information is translated into functional proteins. However, the presence of overlapping reading frames, alternative start codons, and sequence variability complicates the accurate identification of translation initiation start positions. These complexities necessitate the development of sophisticated algorithms capable of deciphering the underlying patterns and signals embedded within these sequences.

In this context, the site of investigation is characterized by the availability of vast genomic data, providing a rich source of information for analysis and prediction. The domain of translation initiation start positions presents an intriguing puzzle due to the interplay between the diverse elements within RNA sequences. The challenges lie not only in deciphering the specific nucleotide combinations that denote start codons but also in understanding the contextual cues surrounding these codons. Moreover, the limitations of existing approaches, which often rely on simplistic algorithms, underscore the need for more advanced methodologies that harness the power of machine learning to improve prediction accuracy.

This thesis seeks to address the general problems faced by the domain by proposing a novel machine-learning algorithm tailored for the prediction of translation initiation start positions. By incorporating multiple features, including RNA sequence information, secondary structure predictions, conservation profiles, and hexamer patterns, the algorithm aims to leverage the comprehensive nature of the data to enhance prediction accuracy. The subsequent chapters will delve into the specifics of the proposed model, its architecture, and its performance evaluation, ultimately contributing to the advancement of knowledge in the field of genomics and RNA sequence analysis.

1.1 Background and Motivation

In this sub-chapter, we will delve into the background and motivation of this thesis, which aims to develop a machine learning algorithm for predicting translation initiation start positions in RNA sequences. Before we delve into the specific details of the algorithm and its architecture, it is important to provide a comprehensive understanding of the scope and challenges associated with this task.

1.1.1 General Description of the Scope

This research focuses on predicting translation initiation start positions within RNA sequences, a critical step in understanding gene expression and protein synthesis. By accurately identifying these start codons, we gain insights into the initiation of protein synthesis and unravel the complexities of cellular processes. The correct prediction of translation initiation sites is essential for deciphering the mechanisms underlying gene regulation and protein functions. Therefore, this thesis aims to develop a sophisticated machine learning algorithm capable of effectively and accurately predicting translation initiation start positions, contributing to advancements in genomics and molecular biology.

1.1.2 Characteristics of the Site

Within the domain of genomic analysis, the prediction of translation initiation start positions presents intriguing opportunities and challenges. The site of investigation encompasses the vast landscape of genomic data available for analysis and prediction. RNA sequences

offer valuable insights into gene expression and protein synthesis, but their inherent complexities pose challenges for accurate start codon identification. Overlapping reading frames, alternative start codons, sequence variability, and contextual cues surrounding start codons further complicate the prediction process. These challenges motivate the development of advanced computational models that can leverage machine learning techniques and integrate multiple informative features to improve prediction accuracy.

Existing approaches in the field often rely on simplistic algorithms that do not fully exploit the potential of machine learning or incorporate comprehensive features. As a result, prediction accuracy may be sub-optimal, and the intricacies of RNA sequences may not be fully captured. Therefore, the motivation behind this thesis is to bridge this gap and propose a novel machine-learning algorithm that can effectively leverage various features, including RNA sequence information, secondary structure predictions, conservation profiles, and hexamer patterns. By integrating these features, we aim to enhance the accuracy of translation initiation start position prediction and provide a deeper understanding of gene expression regulation and protein synthesis.

Through addressing these challenges and pursuing the motivation of improving prediction accuracy and advancing genomic analysis, this thesis aims to contribute to the broader scientific community's knowledge and understanding in the field of genomics and RNA sequence analysis.

1.2 Problem Statement

Predicting translation initiation start positions in RNA sequences poses a significant challenge in the field of genomics and molecular biology. Accurate identification of these start codons is crucial for understanding gene expression and protein synthesis. However, the inherent complexities of RNA sequences, such as overlapping reading frames, alternative start codons, sequence variability, and contextual cues surrounding start codons, make it difficult to reliably predict the true start positions. Existing approaches often rely on simplistic algorithms that do not fully leverage the power of machine learning or incorporate comprehensive features. As a result, prediction accuracy may be sub-optimal, and the intricacies of RNA sequences may not be fully captured.

The significance of accurately predicting translation initiation start positions cannot be

understated. Start codon identification plays a pivotal role in unraveling the mechanisms underlying gene regulation and protein functions. It provides insights into the initiation of protein synthesis and aids in deciphering the complexities of cellular processes. Improved prediction accuracy can lead to a deeper understanding of the factors influencing translation initiation and shed light on the regulation of gene expression.

However, the current state of the art in start codon prediction is limited in several aspects. Existing approaches often lack the sophistication to handle the complexities of RNA sequences and fail to fully exploit the potential of machine learning techniques with all possible features that can be extracted from an RNA sequence. Sub-optimal prediction accuracy and limited integration of diverse features hamper the comprehensive analysis of start codon positions. To overcome these limitations, there is a need for a novel machine learning algorithm specifically designed for predicting translation initiation start positions. Such an algorithm should leverage various features, including RNA sequence information, secondary structure predictions, conservation profiles, and hexamer patterns, to enhance the accuracy and reliability of start codon prediction.

In light of these challenges and limitations, this thesis aims to address the problem of accurately predicting translation initiation start positions in RNA sequences. By developing a sophisticated machine learning algorithm that effectively integrates diverse features and leverages advanced computational models, this research seeks to improve prediction accuracy and deepen our understanding of gene expression and protein synthesis. The subsequent chapters of this thesis will detail the proposed algorithm's architecture, its performance evaluation, and comparisons with existing approaches. Through these efforts, this research endeavors to bridge the gap in start codon prediction and contribute to the advancement of genomic analysis and molecular biology knowledge. a

1.3 Objectives

The primary objectives of this thesis are to develop a machine-learning algorithm for predicting translation initiation start positions in RNA sequences and to advance the field of genomic analysis. These objectives align with the problem statement discussed in the previous sub-chapter and aim to address the limitations of existing approaches while improving prediction accuracy and deepening our understanding of gene expression and protein synthe-

sis.

1. Create a well-annotated dataset: The first objective is to create a well-annotated dataset for training and testing the machine learning algorithm. This dataset will comprise positive examples, including RNA sequences with known translation initiation start positions, and negative examples representing non-start codon regions. The dataset will be carefully curated and annotated, ensuring high-quality and reliable training data for the algorithm. A well-annotated dataset plays a crucial role in machine learning by providing a foundation for the training and evaluation of models. Accurate and comprehensive annotations, which include labeled data points and corresponding ground truth information, enable algorithms to learn patterns, make predictions, and generalize knowledge. Thus, creating this dataset is one of the most crucial objectives of this thesis.

2. Extract informative features: The second objective is to extract informative features from the created dataset that can contribute to better prediction accuracy. This includes analyzing the RNA sequence characteristics, such as nucleotide composition, sequence motifs, and contextual information around start codons. Additionally, features derived from secondary structure predictions, conservation profiles, and hexamer patterns will be extracted to capture relevant biological signals. The extracted features will serve as input to the machine learning algorithm, enhancing its ability to make accurate predictions.

3. Develop a novel machine learning algorithm: The third objective is to develop an innovative machine learning algorithm specifically designed for predicting translation initiation start positions. This algorithm will leverage advanced computational models, such as convolutional neural networks, and integrate diverse features, including RNA sequence information, secondary structure predictions, conservation profiles, and hexamer patterns. The algorithm will be optimized to achieve high prediction accuracy and robustness.

4. Evaluate algorithm performance: The fourth objective is to rigorously evaluate the performance of the developed algorithm using comprehensive datasets. This evaluation will involve assessing the algorithm's accuracy, precision, and recall in predicting translation initiation start positions. The evaluation will provide quantitative measures of the algorithm's effectiveness and serve as a benchmark for comparing with other approaches.

5. Perform hyperparameter tuning: The fifth objective is to perform hyperparameter tuning of the machine learning algorithm. This process involves systematically exploring different combinations of hyperparameters, such as learning rate, batch size, number of layers, and

filter sizes, to optimize the algorithm's performance. Hyperparameter tuning will be conducted using appropriate techniques, such as grid search or random search, to identify the optimal set of hyperparameters that maximize prediction accuracy.

6. Compare with state-of-the-art models: The sixth objective is to compare the performance of the proposed algorithm with existing state-of-the-art models and baseline approaches. By conducting comparative analyses, we will demonstrate the superiority of the developed algorithm in accurately predicting translation initiation start positions. This comparison will highlight the potential of the proposed algorithm and showcase its advancements in the field.

7. Provide comprehensive analysis and interpretation: The seventh objective is to provide a comprehensive analysis and interpretation of the results obtained. This analysis will include discussing the strengths and limitations of the proposed algorithm, identifying potential areas for further improvement, and offering insights into the contribution of different features and hyperparameters to the prediction accuracy. Through this analysis, we aim to contribute to the scientific community's knowledge and understanding in the field of genomics and RNA sequence analysis.

By accomplishing these objectives, this research aims to develop an advanced machine learning algorithm, create a reliable dataset, extract informative features, optimize hyperparameters, and improve the prediction accuracy of translation initiation start positions. These contributions will not only enhance our understanding of gene expression but also advance the field of genomic analysis and enable more accurate predictions in various biological applications.

1.4 Significance of the Study

The significance of this study lies in its potential to advance the field of genomic analysis and improve the accuracy of predicting translation initiation start positions in RNA sequences. By proposing a novel machine learning algorithm and integrating multiple sources of information, including RNA sequence characteristics, secondary structure predictions, conservation profiles, and hexamer patterns, this study aims to provide a comprehensive understanding of gene expression and regulation.

One of the key contributions of this research is the creation of a well-annotated dataset specifically designed for training and testing the machine learning algorithm. The dataset

comprises carefully curated positive and negative examples, derived from UniProt Homo sapiens genome annotation tracks and intergenic regions, respectively. This dataset not only serves as a valuable resource for the development and evaluation of the proposed algorithm but also provides a reliable foundation for future research in the field. The availability of such a dataset enhances the reproducibility of the study and fosters collaboration among researchers working in the domain of genomic analysis.

Furthermore, this study focuses on the extraction of informative features from the dataset to enhance the performance of the algorithm. These features capture various aspects of RNA sequences, including nucleotide composition, sequence motifs, secondary structure predictions, and conservation profiles. By leveraging these features, the algorithm gains a deeper understanding of the underlying biological signals and can make more accurate predictions of translation initiation start positions. The extraction and utilization of informative features contribute to a comprehensive analysis and interpretation of the data, enabling researchers to uncover new insights into gene expression patterns and functional roles of different genes.

The outcomes of this study have broader implications in bio-medicine and biotechnology. The accurate prediction of translation initiation start positions holds great promise for various applications in bio-medicine and biotechnology. By deciphering the precise locations where protein synthesis begins, this research can contribute to a deeper understanding of gene expression mechanisms and provide insights into the functional roles of different genes. In bio-medicine, accurate prediction of translation initiation start positions can have significant implications for the development of therapeutic interventions. Understanding the precise start sites of protein synthesis enables researchers to identify potential drug targets and design interventions that specifically modulate the expression of key genes. This knowledge can lead to the development of targeted therapies aimed at combating various diseases, including genetic disorders, cancers, and neurodegenerative conditions. Moreover, accurate prediction of translation initiation start positions can facilitate the annotation of genomes and the identification of open reading frames (ORFs). By pinpointing the precise locations of protein-coding regions within genomes, this research aids in the accurate annotation and characterization of genes, facilitating genome-wide studies and comparative genomics analyses. This information is crucial for understanding the evolutionary dynamics of genes, identifying conserved regions across species, and uncovering novel genes with potential functional significance.

Moreover, this study makes a notable contribution to the scientific community by propos-

ing a novel machine-learning algorithm and employing rigorous methodologies for dataset construction, feature extraction, and model optimization. The findings and insights obtained from this research have the potential to inspire and guide future studies in genomics, computational biology, and related disciplines. By sharing the methodologies and techniques employed in this study, the research contributes to the collective knowledge and fosters further exploration and advancements in the field of genomic analysis.

In summary, the significance of this study lies in its potential to advance the field of genomic analysis, improve the accuracy of predicting translation initiation start positions, provide a well-annotated dataset, extract informative features, and contribute to the scientific community. The proposed algorithm and methodologies have the potential to revolutionize our understanding of gene expression patterns, uncover new insights into gene function, and drive advancements in bio-medicine and biotechnology. By addressing these crucial aspects, this study aims to make a meaningful impact in the field of genomics and computational biology.

1.5 Translation Initiation Start Positions

By delving into the theoretical aspects of translation initiation start positions, this subchapter provides a comprehensive understanding of the molecular mechanisms underlying protein synthesis initiation. It explores the definition and significance of translation initiation start positions, highlighting the role of start codons and the complexity of the translation initiation process. Additionally, it discusses the regulatory elements and factors that influence translation initiation, shedding light on the intricate regulatory networks that govern gene expression.

Understanding translation initiation start positions has profound implications for various fields, including molecular biology, genetics, and biotechnology. It serves as a foundation for studying gene expression regulation, unraveling the mechanisms that control protein synthesis, and exploring the impact of genetic variations on translation efficiency. Furthermore, accurate prediction of translation initiation start positions contributes to genome annotation efforts, facilitating the identification of protein-coding regions and improving our understanding of gene structures within genomes.

Moreover, the knowledge gained from studying translation initiation start positions has

practical applications in biotechnology and drug discovery. It aids in the design and engineering of synthetic genes for various biotechnological applications, enabling the production of recombinant proteins with precise control over their expression levels. Additionally, it provides insights into potential drug targets and therapeutic interventions by identifying key genes and regulatory elements involved in translation initiation.

1.5.1 Overview of Protein Transcription and Translation Processes

Transcription and translation are essential processes in gene expression, playing a crucial role in the synthesis of proteins. These processes are highly regulated and intricately coordinated to ensure the accurate transfer of genetic information from DNA to functional proteins. Understanding the fundamentals of transcription and translation is vital for comprehending the molecular mechanisms behind protein synthesis and the subsequent prediction of translation initiation start positions.

Transcription is the first step in gene expression, where the genetic information encoded in DNA is transcribed into messenger RNA (mRNA) molecules. This process is carried out by the enzyme RNA polymerase, which binds to the promoter region of the DNA, initiating the synthesis of an mRNA molecule. During transcription, one of the DNA strands acts as a template, and RNA polymerase incorporates complementary RNA nucleotides to form the mRNA molecule. The resulting mRNA undergoes further modifications, such as the addition of a 5' cap and a poly-A tail, as well as the removal of introns, to form a mature mRNA molecule.

Translation, on the other hand, is the process by which the mRNA molecule is translated into a sequence of amino acids, forming a polypeptide chain that ultimately folds into a functional protein. Translation involves three key players: mRNA, ribosomes, and transfer RNA (tRNA). The mRNA molecule carries a sequence of codons, each consisting of three nucleotides, which correspond to specific amino acids. Ribosomes, composed of ribosomal RNA (rRNA) and proteins, serve as the molecular machinery for protein synthesis. They bind to the mRNA and facilitate the decoding of codons by recruiting specific tRNA molecules. Each tRNA molecule carries a specific amino acid and has an anticodon that is complementary to the codon on the mRNA. Through a complex interplay of ribosomes and tRNA molecules, amino acids are added to the growing polypeptide chain in a sequence dictated by the mRNA codons.

Together, transcription and translation govern the central dogma of molecular biology, enabling the transfer of genetic information from DNA to proteins. Understanding these processes provides the foundation for investigating translation initiation start positions and their role in protein synthesis. In the subsequent sections of this thesis, we will delve into the specific aspects of translation initiation, exploring the significance of start codons, the complexity of the initiation process, and the regulatory elements that influence translation initiation.

1.5.2 Definition of Translation Initiation Start Positions

Translation initiation start positions refer to the specific nucleotide positions within an RNA sequence where protein synthesis begins. They mark the site where the ribosome assembles and initiates the process of translation, converting the genetic information encoded in the RNA into a functional protein. These positions are of significant importance in understanding the regulation of gene expression and accurately annotating protein-coding regions within genomes. The determination of translation initiation start positions relies on the identification of start codons, which are typically represented by an AUG (adenine-uracil-guanine) triplet in eukaryotes. AUG serves as the universal start codon, initiating the majority of protein synthesis events. It acts as a signal to recruit the ribosome and initiate the assembly of the translation machinery. However, it is worth noting that alternative start codons, such as GUG and UUG, can also be utilized in specific contexts, leading to the initiation of translation at alternative positions.

1.5.3 Molecular Mechanisms of Translation Initiation

Translation initiation is a complex process involving multiple molecular components and regulatory elements that ensure the accurate initiation of protein synthesis. It can be broadly divided into three main stages: recognition of the start codon, assembly of the translation initiation complex, and ribosome scanning.

The first stage, recognition of the start codon, involves the binding of the small ribosomal subunit, along with initiation factors, to the mRNA sequence. The start codon, typically the AUG triplet, serves as the point of recognition. The anticodon of the initiator tRNA (tRNAiMet) forms a base pairing interaction with the start codon, facilitated by initiation factors. This recognition event ensures the proper alignment of the ribosomal sub-units and the positioning of the initiator tRNA at the start codon.

In the second stage, the assembly of the translation initiation complex takes place. Additional initiation factors are recruited to the complex, along with the large ribosomal subunit. These initiation factors play crucial roles in stabilizing the interaction between the mRNA, tRNAiMet, and ribosomal subunits. They assist in positioning the components of the translation machinery in a manner that enables efficient and accurate initiation of protein synthesis. Once the translation initiation complex is fully assembled, the ribosome proceeds to the third stage, which is ribosome scanning. During this process, the ribosome traverses along the mRNA molecule, searching for the correct start codon. The scanning mechanism relies on the recognition of specific nucleotide sequences that guide the ribosome to the start codon. In eukaryotes, a notable sequence is the Kozak consensus sequence, which surrounds the start codon. The Kozak consensus sequence aids in proper positioning and increases the efficiency of translation initiation by providing additional signals for ribosome recognition.

1.5.4 Regulatory Elements and Factors Influencing Translation Initiation

Translation initiation is a finely regulated process influenced by a range of regulatory elements and factors that modulate the efficiency and specificity of protein synthesis. These elements include upstream open reading frames (uORFs), regulatory RNA structures, and RNA-binding proteins, each playing a crucial role in controlling translation initiation.

uORFs are short coding sequences that reside upstream of the main coding region within an mRNA molecule. They have the potential to influence translation initiation by various mechanisms. For instance, uORFs can impact the availability of initiation factors by sequestering them or competing for their binding, leading to altered initiation rates. Additionally, uORFs can function as alternative start sites, diverting the ribosome's attention and affecting the translation of downstream proteins. The presence and characteristics of uORFs within an mRNA sequence can significantly impact the regulation of gene expression and protein synthesis.

Regulatory RNA structures also contribute to the control of translation initiation. These structures can form within the mRNA molecule, either around the start codon or in other regions, and influence the accessibility of the start codon or the binding of initiation factors and ribosomes. Secondary structures, such as stem-loops, hairpins, or pseudoknots, can hinder or facilitate ribosome scanning, thereby affecting translation initiation efficiency. RNA motifs,

including internal ribosome entry sites (IRES), play a unique role by bypassing the traditional scanning mechanism and enabling cap-independent translation initiation.

In addition to structural elements, RNA-binding proteins play a vital role in modulating translation initiation. These proteins interact with specific mRNA sequences or structures, exerting regulatory control over protein synthesis. RNA-binding proteins can act as translational activators, promoting translation initiation by facilitating ribosome recruitment or stabilizing the initiation complex. Conversely, they can function as translational repressors, inhibiting translation initiation through various mechanisms, such as preventing ribosome assembly or promoting mRNA degradation. The interplay between RNA-binding proteins and mRNA molecules adds another layer of complexity to the regulation of translation initiation, influencing gene expression dynamics in response to cellular signals and environmental cues.

1.6 Importance of Predicting TIS

Accurate prediction of TIS positions holds significant importance in various domains of biological research and applications. The precise identification of these positions provides valuable insights into gene expression regulation, aids in genome annotation, supports functional genomics studies, accelerates drug discovery efforts, enhances genome functional annotation, and enables advancements in synthetic biology and biotechnology. The importance of predicting TIS positions can be further elucidated through the following points:

1. Understanding of Gene Expression Regulation: Prediction of TIS positions contributes to a comprehensive understanding of gene expression regulation. By accurately identifying the specific locations where translation initiates, researchers can unravel the intricate mechanisms governing protein synthesis and its control within cells. This knowledge is fundamental for deciphering complex cellular processes, understanding developmental pathways, and unraveling the molecular basis of diseases.

2. Identification and Annotation of Protein-Coding Regions: Accurate prediction of TIS positions aids in the identification and annotation of protein-coding regions within genomes. By precisely delineating the boundaries of coding regions, researchers can refine gene models and improve genome annotations. Reliable prediction of TIS positions enhances our understanding of the functional elements encoded in the genome and provides valuable information for comparative genomics studies and evolutionary analyses.

3. Facilitation of Functional Genomics Studies: Prediction of TIS positions plays a vital role in functional genomics studies. It assists in prioritizing candidate genes for further experimental investigation, allowing researchers to focus their efforts on protein-coding regions with potential functional significance. By accurately predicting TIS positions, researchers can identify genes involved in specific biological processes, pathways, or diseases, providing valuable guidance for targeted experimental studies.

4. Acceleration of Drug Discovery and Development: Accurate prediction of TIS positions contributes to the process of drug discovery and development. By identifying the precise start sites of protein synthesis, researchers gain insights into the molecular targets of therapeutic relevance. This information is crucial for designing drugs that specifically target disease-associated proteins or modulate their expression levels. Predicting TIS positions facilitates the development of more effective and targeted therapeutic interventions.

5. Enhancement of Genome Functional Annotation: Prediction of TIS positions contributes to the comprehensive functional annotation of genomes. By accurately identifying TIS positions, researchers can assign functional annotations to protein-coding regions, thereby enhancing our understanding of gene function and diversity. This information supports comparative genomics studies, evolutionary analyses, and the exploration of gene families and protein domains.

6. Support for Synthetic Biology and Biotechnology Applications:

Accurate prediction of TIS positions has implications in synthetic biology and biotechnology. It enables the rational design and engineering of synthetic genetic circuits and pathways by precisely controlling the initiation of protein synthesis. Predicting TIS positions aids in optimizing protein expression levels, fine-tuning gene expression dynamics, and improving the efficiency of recombinant protein production systems.

1.7 Existing Approaches and Limitations

Support Vector Machines (SVMs) have been a prominent tool in the prediction of Translation Initiation Start Positions (TIS) within RNA sequences. Originally designed for binary classification tasks, SVMs offer an efficient approach to identifying TIS codons. They are adept at finding decision boundaries that maximize class separation in RNA sequences, making them particularly suited for this task. However, it's important to note that SVMs, while

effective, have inherent limitations. Their simplicity compared to more complex models like artificial neural networks (ANNs) means they may struggle to capture highly intricate patterns within the data. This limitation raises questions about their ability to match the predictive performance of more intricate models when facing datasets with nuanced features.

Then, some early neural network models explored the potential of artificial neural networks (ANNs) with relatively simple architectures, often consisting of only a single hidden layer. These models aimed to harness the power of neural networks to identify TIS codons within RNA sequences. While they marked a significant departure from traditional machine learning methods, their shallow architectures, and limited complexity posed certain challenges. The use of single-layer neural networks could potentially result in difficulties when attempting to capture intricate and non-linear patterns within RNA sequences. This limitation, combined with the evolving nature of TIS prediction, raises questions about the ability of these early neural network models to adapt to more complex datasets and outperform modern, deeper architectures.

In recent years, the field of TIS prediction has witnessed a significant shift with the introduction of modern neural network architectures, primarily leveraging Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models hold the potential to revolutionize TIS prediction in several ways. Hierarchical Feature Extraction: CNNs excel at capturing local patterns within RNA sequences, including motifs and secondary structures. Their hierarchical feature extraction abilities allow them to learn complex representations directly from the data, which is critical for precise TIS identification. Sequence Context: RNNs, with their ability to model sequential dependencies, can capture the contextual information within RNA sequences. This understanding of nucleotide order and relationships is fundamental for distinguishing TIS codons from the surrounding genomic sequences. However, it's imperative to address a significant concern associated with these modern approaches. Many of these models, despite showcasing remarkable results, suffer from a severe lack of reproducibility. They often provide limited or no information regarding their dataset sources or construction, making it nearly impossible for other researchers to replicate their findings. This absence of transparency raises questions about the reliability and generalizability of their results. Moreover, the reluctance to share code further exacerbates the issue of non-reproducibility. Without access to the model code, the scientific community is left unable to scrutinize the inner workings and assumptions of these approaches.

1.8 Summary

In conclusion, the first chapter has laid the foundation for the exploration into the development of a Machine Learning Algorithm for Predicting Translation Initiation Start Positions (TIS). As this thesis unfolds, the following chapters will present:

Chapter 2: Dataset Construction

Chapter 2 provides a detailed account of data collection and preprocessing efforts, explaining the curation of the dataset, comprising positive sets sourced from Start Codon Annotations and negative sets derived from Non-Start Codon Annotations.

Chapter 3: Model Architecture

Chapter 3 serves as the core of the thesis, offering an in-depth view of the proposed model. It consists of multiple branches, each specializing in different aspects of TIS prediction. The design choices, architectural details, and performance evaluation of each branch are elucidated.

Chapter 4: Experimental Results and Analysis

Chapter 4 presents the results obtained from the model. Evaluation metrics are introduced to assess performance and make comparisons with baseline and state-of-the-art models. This chapter concludes with a comprehensive analysis of the results, highlighting the strengths and limitations of the approach.

Chapter 5: Conclusion and Future Work

The final chapter of the thesis, Chapter 5, encapsulates the contributions and summarizes key findings. It acknowledges the limitations of the model and proposes avenues for future research, highlighting potential enhancements and directions to further advance the field of TIS prediction.

As these chapters unfold, the aim is to provide a holistic understanding of the machine learning methodology employed for TIS prediction. This thesis will try to underscore the significance of this work within the realm of genomics and the exciting prospects it holds for future developments in this critical domain.

Chapter 2

Dataset

2.1 Introduction

In this chapter, the process of constructing the dataset used in the study to predict translation initiation start positions is presented. A well-curated dataset forms the foundation for training an accurate and robust machine-learning model. In this section, an overview of the dataset used is provided, and emphasized the significance of a high-quality dataset in the context of predicting translation initiation start positions.

2.1.1 Overview of the Dataset

The dataset employed in this study comprises RNA sequences derived from the Swiss-Prot database, specifically the UniProt Homo sapiens genome annotation tracks. These annotation tracks represent the mapping of a species' reference proteome and sequence annotations to its reference genome. Utilizing this rich source of genomic information, it is aimed to construct a comprehensive dataset that captures the diversity of translation initiation start positions in Homo sapiens.

In addition to the positive set obtained from the UniProt database, a carefully constructed negative set was incorporated to enhance the model's training. The negative set was constructed by utilizing the Ensembl Genome Annotation (GTF) file for Homo sapiens. This GTF file contains various attributes associated with each feature, including gene-related information, transcript-related information, exon-related information, and protein-related information.

2.1.2 Significance of a Well-Curated Dataset

Having a well-curated dataset plays a pivotal role in training a reliable machine-learning model. In the context of predicting translation initiation start positions, the quality and comprehensiveness of the dataset directly impact the model's performance and generalizability. A carefully constructed dataset enables the model to learn the intricate patterns and features associated with translation initiation start positions, thereby enhancing its ability to accurately predict these crucial sites.

By utilizing a well-curated dataset, several key advantages can be addressed. Firstly, it allows the capture of a wide range of start codon positions and their associated genomic contexts, ensuring that the model is exposed to diverse scenarios. This helps the model generalize well to unseen sequences and improves its ability to accurately predict translation initiation start positions in real-world scenarios.

Furthermore, a high-quality dataset enables the uncovering of valuable insights into the regulatory mechanisms underlying translation initiation. By analyzing the dataset, sequence motifs can be identified, secondary structures, or other features that are strongly associated with translation initiation start positions. This knowledge can contribute to a deeper understanding of gene expression regulation and aid in the development of novel therapeutic interventions targeting translation initiation.

Overall, the significance of a well-curated dataset cannot be overstated. It forms the backbone of the study, providing the necessary training examples to develop a robust machine-learning model for predicting translation initiation start positions. In the subsequent sections, a delve into the details of data collection, preprocessing steps, and the construction of the positive and negative sets, which collectively contribute to the establishment of a reliable and representative dataset for our research.

Moreover, the construction of a well-curated dataset encompasses the assembly of both positive and negative sets. The positive set comprises well-annotated start codon positions, while the negative set contains regions that do not correspond to true start codons. The inclusion of a comprehensive negative set is particularly crucial as it allows the model to discriminate between genuine start codons and non-start codon sequences accurately. By incorporating a diverse range of non-start codon regions, the model can learn the discriminating features associated with translation initiation, thereby improving its predictive capabilities.

To ensure the quality and generalizability of the negative set, several considerations were

taken into account during its construction. Careful selection of genomic regions, covering various regions across the genome, helps mitigate potential biases and ensures a representative dataset. Additionally, stringent filtering criteria were applied to minimize the inclusion of regions that might inadvertently resemble start codons. By addressing these factors, the negative set is designed to challenge the model and enhance its ability to distinguish between true start codons and non-start codon sequences in a robust and unbiased manner.

By incorporating a well-curated dataset, encompassing both positive and negative sets, this research aims to train a machine-learning model capable of accurately predicting translation initiation start positions. The subsequent sections delve into the details of the data collection process, including the acquisition of the positive set from the UniProt/SwissProt database and the construction of the negative set using the Ensembl GTF file. These steps collectively contribute to the establishment of a reliable and representative dataset, enabling the model to learn the intricate patterns and features associated with translation initiation start positions.

2.2 Data Collection and Preprocessing

2.2.1 Positive Set: Start Codon Annotations

The positive dataset for this thesis was gathered from the UniProt/SwissProt database. Specifically, the data were found from this ftp folder, and in the UP000005640_9606 signal .bed file. This .bed file contains a bed representation of signal sequence annotations within the species, and the species is the Homo sapiens, which is denoted by the UP000005640_9606.

A .bed file is a file format used to represent genomic intervals, such as regions of a genome that correspond to genes, regulatory elements, or other functional features.s

The name ”bed” stands for ”Browser Extensible Data,” as the format was originally developed for the UCSC Genome Browser, which allows users to visualize genomic data in a web browser.

A .bed file consists of tab-delimited fields that specify the chromosome, start and end coordinates, and optionally a name and additional metadata for each interval. The format can also be extended to include additional columns for other types of data, such as scores or strand orientation.

The format for the .bed file is:

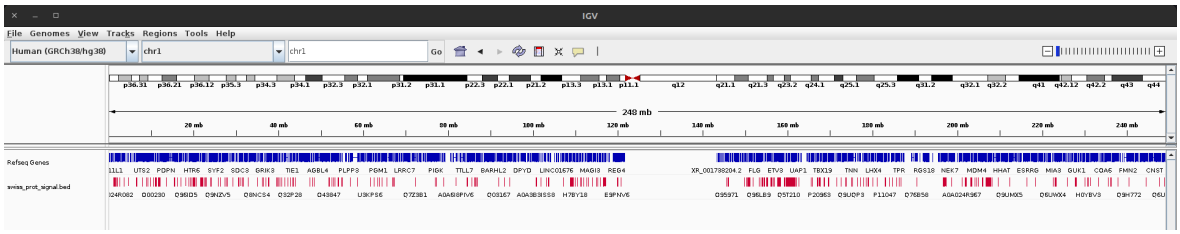

Figure 2.1: .bed file opened in the IGV Browser

```
chromosome  start   end     name    score   strand      other inform
```

Where:

- chromosome: the name of the chromosome or contig (e.g. chr1, chrX, contig123)
- start: the starting position (0-based) of the interval on the chromosome
- end: the ending position (1-based, inclusive) of the interval on the chromosome
- end: the ending position (1-based, inclusive) of the interval on the chromosome
- name: an optional name or identifier for the interval
- score: an optional score or value for the interval
- strand: the strand orientation of the interval, either ”+” for the forward strand or ”-” for the reverse strand. If the strand information is not available, a ”.” is used instead.

An example of a line of the bed file for the study is:

```
chr1 1091489 1091543 A0A024R082 0 -1091489 1091543 204,0,51
1 54 0 . M1-S18; signal peptide_1-18
```

A .bed file can be used for a variety of genomic analyses, including identifying genomic regions that are deferentially expressed, determining chromatin accessibility, and detecting genetic variants.

After downloading the .bed file, the next step in the data preprocessing phase involved merging the overlapping peaks. This was done to ensure that the resulting dataset contained non-overlapping peaks with consistent information. The merging process combined intervals that shared overlapping regions into a single peak, while retaining the original name, score, and strand information.

The merged file now consists of non-overlapping peaks, where each peak represents a unique region of interest. The start and end coordinates of the peaks have been adjusted to eliminate any overlap, ensuring that each peak represents a distinct genomic region. By performing this merging step, a refined dataset was obtained that is more suitable for subsequent analysis and modeling.

In addition to downloading the .bed file from the UniProt/SwissProt database, another important source of data for this study was a .bed file obtained from the Biomart database. The Biomart database provides a wide range of genomic and proteomic data, allowing researchers to retrieve specific annotations and features related to various organisms.

The purpose of obtaining the .bed file from Biomart was to obtain additional genomic annotations that would complement the data obtained from the UniProt/SwissProt database. By intersecting the two datasets, it was aimed to ensure the inclusion of well-annotated data in the study.

The intersection of the .bed files involved identifying the common regions between the datasets obtained from UniProt/SwissProt and Biomart. This process allowed us to select and retain the genomic regions that were consistently annotated in both datasets, ensuring a comprehensive and well-annotated dataset for our analysis.

By incorporating the additional annotations from Biomart, we enriched the dataset with valuable information about the genomic features associated with the translation initiation start positions. This integration of data sources enhanced the accuracy and reliability of our model by incorporating multiple layers of annotation and increasing the overall understanding of the genomic context.

Then, each entry in the resulted intersected .bed file is modified, by adding 2Kbp before the Start and 10Kbp after the Start, creating this way a 12Kbp sequence for each entry, where the start of the protein is after 2000 bp. So the TIS position now is in the 2000 place. (why do we do that ?)

To retrieve the sequence information for each of the genomic regions specified in the .bed files, the reference genome fasta file was utilized. This fasta file, obtained from the UCSC Genome Browser, serves as a comprehensive resource containing the genetic sequence of the reference genome.

By extracting the sequence corresponding to the genomic intervals specified in the .bed files, it was ensured that the dataset included the actual nucleotide sequences associated with

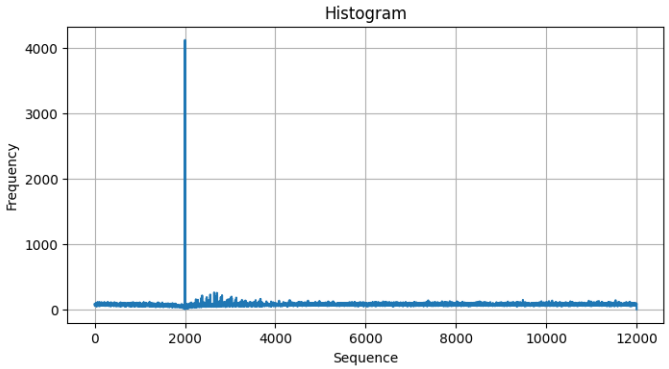

Figure 2.2: AUG position count

the translation initiation start positions. This sequence of information played a crucial role in training and evaluating the machine learning model, as it provided the necessary input data for the prediction algorithms.

The use of the reference genome fasta file from the UCSC Genome Browser facilitated the accurate alignment of the genomic coordinates with their corresponding nucleotide sequences, enabling a comprehensive analysis of the translation initiation start positions within the genomic context.

Having the sequences for each input, a count of the AUG in the sequence was done, to check how many AUG are in a sequence.

As it can be seen from figure 2.2, most of the AUG codons are in the 2000^{th} position, where they were placed, but also it can be observed that lots of other AUG can be found, all around the sequence.

2.2.2 Negative Set: Non-Start Codon Annotations

The negative set was constructed by obtaining an Ensembl .gtf file for homo-sapiens, which contains comprehensive genomic feature information. This file was then transformed into the .bed format for a simplified representation of genomic intervals.

To create the negative set, the entries in the .bed file were merged if they were within a maximum distance of 2500 base pairs (bps), ensuring that overlapping or closely located regions were consolidated. The start positions of the merged entries were adjusted by subtracting 1000 bps, and 1000 bps were added to the end positions to expand the feature information beyond well-annotated regions.

The modified entries were then analyzed to identify complementary regions located far

away from any well-annotated regions. These regions, devoid of start codons, formed the negative set for the study. This approach ensured that the negative set provided a suitable contrast to the positive set and contributed to the balanced dataset used for training and evaluating the machine learning model.

The resulting negative set, comprising non-start codon regions, played a crucial role in the accurate prediction of translation initiation start positions. It served as an essential reference for the model to differentiate between true start codon regions and other genomic regions, enhancing the model's prediction capabilities and overall performance.

For each entry in the negative set, the process of constructing a new entry involved extracting all ATG codons from the original region. The ATG codon serves as the start codon for protein translation. To ensure a comprehensive dataset, a fixed distance of 310 bps was considered from each ATG codon.

By applying this distance criterion, a sequence window of 303 bps (corresponding to the length required by the model) was established for each ATG codon. These individual sequences were then merged to form a massive dataset per chromosome, encompassing all potential start codon regions within the specified distance.

The final output file consisted of a tab-separated format, providing information about the chromosome, start and end positions of the genomic interval, and the corresponding sequence of 303 bps. Each entry in this file represented a potential translation initiation start position and served as valuable training data for the machine learning model.

This process ensured that the dataset encompassed a wide range of potential start codon regions, enabling the model to learn and make predictions based on diverse genomic contexts. By including multiple entries per chromosome, the dataset accounted for the variability in translation initiation start positions across different genomic regions.

The construction of each new entry in this manner facilitated the development of a comprehensive and representative negative set, supporting the model's ability to distinguish true start codons from non-start codon regions.

Chapter 3

Model Architecture

3.1 Overview of the Proposed Model

The proposed model is designed to accurately predict translation initiation start positions by leveraging a multi-branch architecture. This approach capitalizes on the advantages offered by combining multiple sources of information to enhance prediction accuracy.

The model comprises several key components, each contributing to the prediction process. These components include the RNA sequence branch, the Contrafold branch, the conservation branch, and the hexamer branch. Each branch focuses on extracting specific features from the input data to capture relevant patterns and signals associated with translation initiation start positions.

In the RNA sequence branch, the input data consists of the RNA sequences themselves. To extract meaningful features, we employ a convolutional neural network (CNN) architecture that processes subsequences of 23 base pairs (bps). These subsequences are represented as one-hot vectors to capture the underlying patterns that influence the presence of a start codon.

The Contrafold branch utilizes the Contrafold algorithm to predict the secondary structure of the RNA sequences. A parallel CNN is employed to extract features from the predicted secondary structures. This branch provides complementary information about the structural characteristics of the RNA sequences, which can be indicative of translation initiation sites.

The conservation branch takes advantage of the conservation scores of the RNA sequences. These scores reflect the level of evolutionary conservation at each position, highlighting potentially functional regions. By utilizing another parallel CNN, the conservation branch extracts features related to the evolutionary conservation of the RNA sequences, aid-

ing in the identification of translation initiation start positions.

The hexamer branch considers a larger context by incorporating the 150 base pairs (bps) preceding and following the start codon. This extended region is segmented into hexamer strings, resulting in a substantial matrix of hexamer representations. Another parallel CNN is then applied to extract features from these hexamer strings, capturing higher-level patterns and long-range dependencies that may influence translation initiation.

The features extracted from each branch are integrated to make the final prediction. By combining information from multiple sources, the proposed model aims to capture a more comprehensive representation of the RNA sequences, leading to improved accuracy in predicting translation initiation start positions.

In the subsequent subchapters, an examination of the specific architectures, design choices, performance evaluations, and integration strategies employed within each branch. Combining these branches aims to provide a robust and accurate model for predicting translation initiation start positions, contributing to advancements in gene annotation and functional genomics.

3.1.1 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in machine learning model development, involving the systematic exploration of different values for hyperparameters to identify the optimal configuration. Hyperparameters, such as learning rate, batch size, number of filters, kernel size, and dropout rate, significantly impact the model's performance and generalization capabilities.

In this study, hyperparameter tuning was performed across all the proposed branches (except the hexamer branch) which will be analyzed later in the thesis. A systematic search was conducted, considering different combinations of hyperparameter values, to identify the optimal settings that yield the best performance on the validation set. This iterative process helps to fine-tune the model and enhance its predictive capabilities.

By performing hyperparameter tuning, it was ensured that all the models were configured optimally and capable of capturing the most relevant features from the input data. The results presented later in this thesis are based on the models' performance after the hyperparameter tuning process.

3.1.2 Metrics

To evaluate the performance of the different architectures in predicting translation initiation start positions, several commonly used evaluation metrics are employed. These metrics provide quantitative measures of the models' accuracy, precision, and recall, enabling a comprehensive assessment of their predictive capabilities.

Accuracy: Accuracy measures the overall correctness of the predictions made by the model. It is calculated as the ratio of the total number of correct predictions to the total number of predictions. A higher accuracy score indicates a higher level of correctness in the model's predictions.

Precision: Precision quantifies the proportion of true positive predictions out of all positive predictions made by the model. It is computed as the ratio of true positives to the sum of true positives and false positives. Precision provides insights into the model's ability to avoid false positives and make precise predictions.

Recall (Sensitivity): Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances correctly identified by the model. It is calculated as the ratio of true positives to the sum of true positives and false negatives. Recall reflects the model's ability to detect true positives and is particularly important when the identification of positive instances is critical.

These evaluation metrics collectively offer a comprehensive understanding of the RNA Sequence Branch's performance in predicting translation initiation start positions. By examining accuracy, precision, and recall, we can gain insights into the model's overall correctness, precision in positive predictions, and sensitivity in identifying true positive instances. These metrics will be used to analyze and assess the performance of the RNA Sequence Branch throughout the training process and compare it with other models and baselines

3.2 RNA Sequence Branch

3.2.1 Architecture and Design Choices

The RNA Sequence Branch plays a crucial role in the overall model architecture, focusing on extracting relevant features from the RNA sequences to predict translation initiation start positions. By leveraging convolutional neural networks (CNNs), this branch aims to capture

local patterns and sequence motifs that are indicative of the presence or absence of start codons.

RNA sequences contain vital information about the genetic code and play a fundamental role in protein synthesis. Analyzing these sequences can provide insights into the initiation of translation, a key step in protein production. By identifying the exact location of translation initiation start sites, the understanding of gene expression and protein function can be enhanced.

In this branch, the RNA sequences are preprocessed and encoded as one-hot vectors, allowing the model to capture the sequence composition and local patterns. The use of a 23-base pair (bp) window size was chosen to consider a sufficient context around the translation initiation start codon (AUG), as well as additional bases before and after it. This window size aligns with known biological characteristics, as the AUG codon is the most common start codon and is typically surrounded by specific nucleotide sequences.

3.2.2 Architecture and Design Choices

The architecture of the RNA Sequence Branch consists of multiple layers, including convolutional and pooling layers, which are designed to capture hierarchical features at various scales. The specific configuration of these layers, such as kernel sizes, number of filters, and activation functions, is carefully chosen to optimize the model's ability to capture start codon-related patterns.

In more detail, the RNA Sequence Branch consists of three Conv1D layers with max-pooling layers in between. These layers form the core architecture for capturing local patterns and relevant features associated with translation initiation start positions.

The Conv1D layers are designed to extract features from the encoded RNA sequences by applying a set of learnable filters across the sequence. Following each Conv1D layer, a max pooling layer is applied to downsample the feature maps, retaining the most important information while reducing spatial dimensions.

To improve generalization and mitigate overfitting, dropout layers are inserted after each max pooling layer. Dropout randomly deactivates a fraction of neurons during training, preventing over-reliance on specific features and encouraging the model to learn more robust representations.

The final output from the Conv1D layers is passed through a flatten layer, which trans-

forms the multidimensional feature maps into a flattened vector representation. This vector can be fed into subsequent fully connected layers for further processing and prediction.

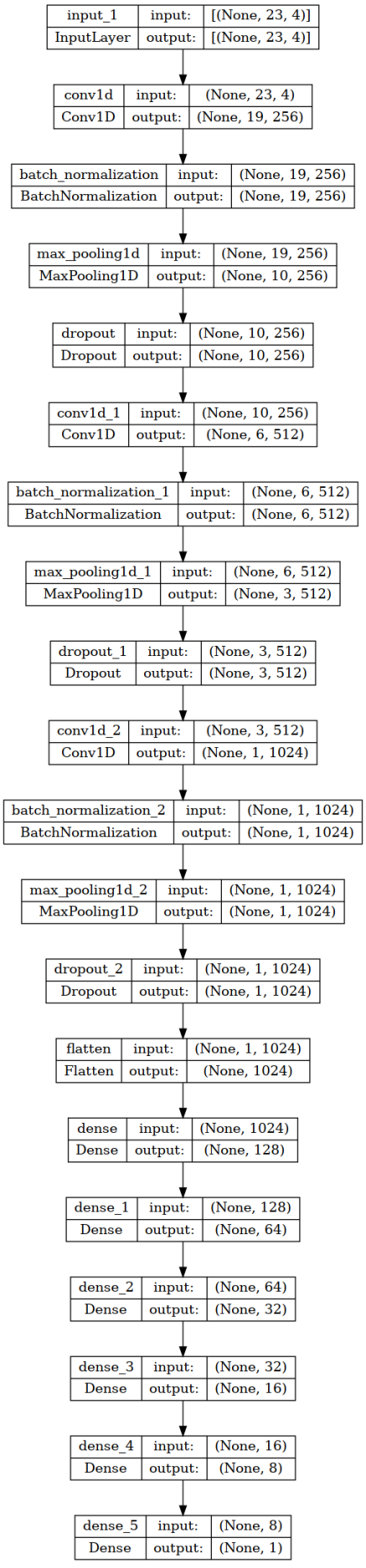

Figure 3.1: Sequence Model Architecture

3.2.3 Performance Evaluation

To ensure optimal performance of the RNA Sequence Branch, hyperparameter tuning was performed to identify the best configuration for the model. The hyperparameters considered in the tuning process included the learning rate and the learning algorithm, batch size, number of filters, kernel size, activation functions, and Dropout rates. Through a systematic search of various combinations of hyperparameter values, the optimal settings were identified based on their performance on the validation set. The best hyperparameters determined for the RNA Sequence Branch were:

- Learning rate: 0.01
- Learning algorithm: Adam
- Batch size: 64
- Filters: 256, 512, 1024, respectively for each CNN layer
- Kernel sizes: 5, 5, 3, respectively for each CNN layer
- Activation Functions: relu, relu, relu
- Dropout rates: 0.2, 0.2, 0.1, respectively for each Dropout layer

The results from the RNA Sequence branch were :

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.3862122	0.83333	0.846062	0.814942	709	129	741	161

3.3 Contrafold Branch

The Contrafold Branch is a crucial component of the model architecture, aiming to incorporate secondary structure information from RNA sequences into the prediction of translation initiation start positions. To achieve this, the branch leverages the first edition of Stanford's Contrafold code, which predicts the secondary structure of an RNA sequence and produces a sequence of characters representing base pairings ('(', ')') and unpaired positions ('.'). This secondary structure prediction offers valuable insights into the potential folding patterns and structural constraints within the RNA sequences.

In the Contrafold Branch, the output of the Contrafold code is preprocessed by converting the characters ('(', ')', and '.') into a numerical representation using one-hot encoding. This vectorization step transforms the secondary structure sequence into a binary matrix, which can be readily fed into the subsequent convolutional layers of the model.

The architecture of the Contrafold Branch shares similarities with the RNA Sequence Branch, comprising convolutional layers and pooling layers. However, instead of directly processing the RNA sequences, the Contrafold Branch takes the one-hot encoded secondary structure representation as input. The convolutional layers within this branch are specifically designed to extract structural features and capture patterns within the secondary structure sequence. Through experimentation and optimization, the configuration of these convolutional layers, including kernel sizes, number of filters, and activation functions, is carefully determined to enhance the model's ability to learn meaningful structural representations.

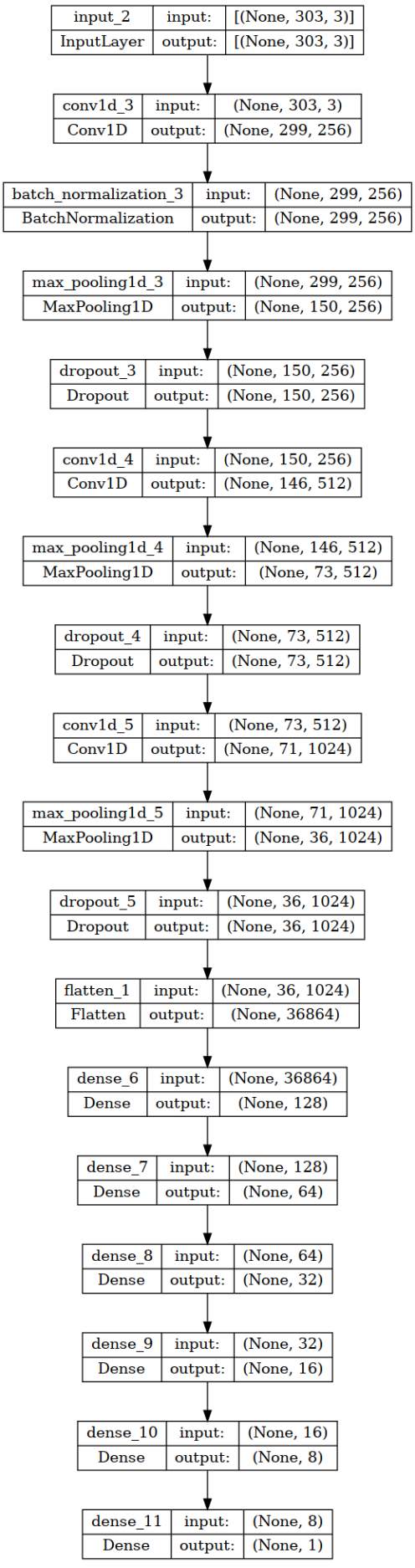

Figure 3.2: Contrafold Model Architecture

3.3.1 Architecture and Design Choices

To ensure optimal performance of the Contrafold Branch, hyperparameter tuning was performed to identify the best configuration for the model. The hyperparameters considered in the tuning process included the learning rate and the learning algorithm, batch size, number of filters, kernel size, activation functions, and Dropout rates. Through a systematic search of various combinations of hyperparameter values, the optimal settings were identified based on their performance on the validation set. The best hyperparameters determined for the Contrafold Branch were:

- Learning rate: 0.0001
- Learning algorithm: Adam
- Batch size: 64
- Filters: 128, 1024, 1024, respectively for each CNN layer
- Kernel sizes: 5, 5, 3, respectively for each CNN layer
- Activation Functions: relu, selu, selu
- Dropout rates: 0.2, 0.1, 0.1, respectively for each Dropout layer

The results from the Contrafold branch were :

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.536843	0.73103	0.68075	0.870114	757	355	515	113

3.4 Conservation Branch

The Conservation Branch plays a vital role in incorporating information about the conservation levels of genomic regions surrounding translation initiation start positions. Conservation scores provide valuable insights into the evolutionary importance and functional relevance of specific genomic regions.

To capture the conservation information, the Conservation Branch utilizes the same 23 base pairs as the RNA Sequence Branch, which include the 15 base pairs before the translation initiation start (TIS) codon, the TIS codon itself, and an additional 5 base pairs. This

configuration ensures that the conservation information aligns with the corresponding RNA sequence.

To obtain the conservation scores, the constructed input, along with the corresponding chromosome information, is passed to the pyBigWig Python library. This library enables the retrieval of conservation profiles from a precomputed bigWig file, which contains the conservation scores across the genome. By passing the file to the pyBigWig library, an array of conservation values is obtained, corresponding to the specified genomic region of interest.

The conservation scores reflect the level of evolutionary conservation for each position within the region. Higher conservation scores indicate positions that have remained conserved across different species and are likely to have functional importance.

To further improve the performance of the model, a scaler is trained on the training data using the conservation scores. This scaler ensures that the conservation values are appropriately scaled and consistent across the training and test datasets. By applying the scaler to the test data as well, the model can effectively handle conservation scores from unseen genomic regions.

Once the preprocessing steps are completed, the final input for the Conservation Branch is obtained, which incorporates both the RNA sequence and the corresponding conservation scores. This joint input allows the model to capture the relationship between the sequence information and the evolutionary conservation of the genomic region.

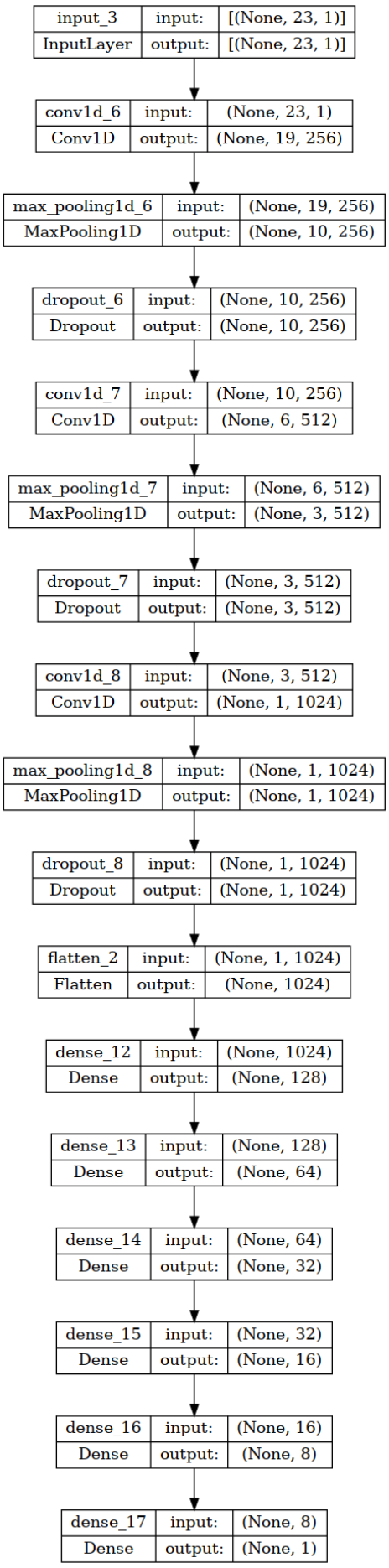

Figure 3.3: Conservation Model Architecture

3.4.1 Architecture and Design Choices

To ensure optimal performance of the Conservation Branch, hyperparameter tuning was performed to identify the best configuration for the model. The hyperparameters considered in the tuning process included the learning rate and the learning algorithm, batch size, number of filters, kernel size, activation functions, and Dropout rates. Through a systematic search of various combinations of hyperparameter values, the optimal settings were identified based on their performance on the validation set. The best hyperparameters determined for the Conservation Branch were:

- Learning rate: 0.0001
- Learning algorithm: Adam
- Batch size: 64
- Filters: 128, 1024, 1024, respectively for each CNN layer
- Kernel sizes: 5, 5, 3, respectively for each CNN layer
- Activation Functions: relu, selu, selu
- Dropout rates: 0.2, 0.1, 0.1, respectively for each Dropout layer

The results from the Conservation branch were :

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.44620147	0.776436	0.904201	0.618390	538	57	813	332

3.5 Hexamer Branch

The Hexamer Branch of the proposed model incorporates the hexamer feature, which provides additional information about the presence and distribution of specific hexamer patterns in the genomic regions surrounding translation initiation start (TIS) codons. Hexamers are six-base sequences that may contain valuable information related to translation initiation and gene regulation.

To construct the input for the Hexamer Branch, 149 base pairs before the TIS codon and 149 base pairs after the TIS codon are selected, resulting in a total of 301 base pairs. This configuration ensures that the hexamer patterns capture the relevant context around the TIS

codon, allowing the model to capture the patterns and relationships among different hexamer sequences.

From the selected region, a hexamer matrix is constructed. Each position in the matrix corresponds to a hexamer sequence, and the value in that position represents the number of times that hexamer has appeared in the sequence. If a particular hexamer is not present in the selected region, a value of 0 is assigned to that position in the matrix. This matrix provides a representation of the hexamer patterns within the genomic region.

To enhance the performance of the model, preprocessing steps are applied to the hexamer matrix, including scaling. Similar to the Conservation Branch, a scaler is trained on the training data using the hexamer matrix. This scaler ensures that the hexamer values are appropriately scaled and consistent across the training and test datasets.

Once the scaler is trained, it is applied to both the training and test data to transform the hexamer matrix. This step ensures that the input provided to the network maintains the same scaling and preserves the relationships within the hexamer patterns. By incorporating the scaled hexamer matrix as input, the Hexamer Branch allows the model to learn and capture the patterns and relationships among different hexamer sequences in the genomic regions surrounding the TIS codons.

3.5.1 Architecture and Design Choices

In the case of the Hexamer Branch, it's important to note that due to the computational complexity and time constraints involved in training and evaluating each model, a comprehensive hyperparameter tuning process was not performed. Given the extensive nature of the hexamer feature extraction and the resulting large matrix, training and evaluating multiple model configurations with different hyperparameter settings would have required a significant amount of computational resources and time. Therefore, a decision was made to prioritize model development and evaluation using a set of reasonable default hyperparameters that are commonly found to be effective in similar tasks. Thus, the same options that were used for the RNA sequence Branch were used.

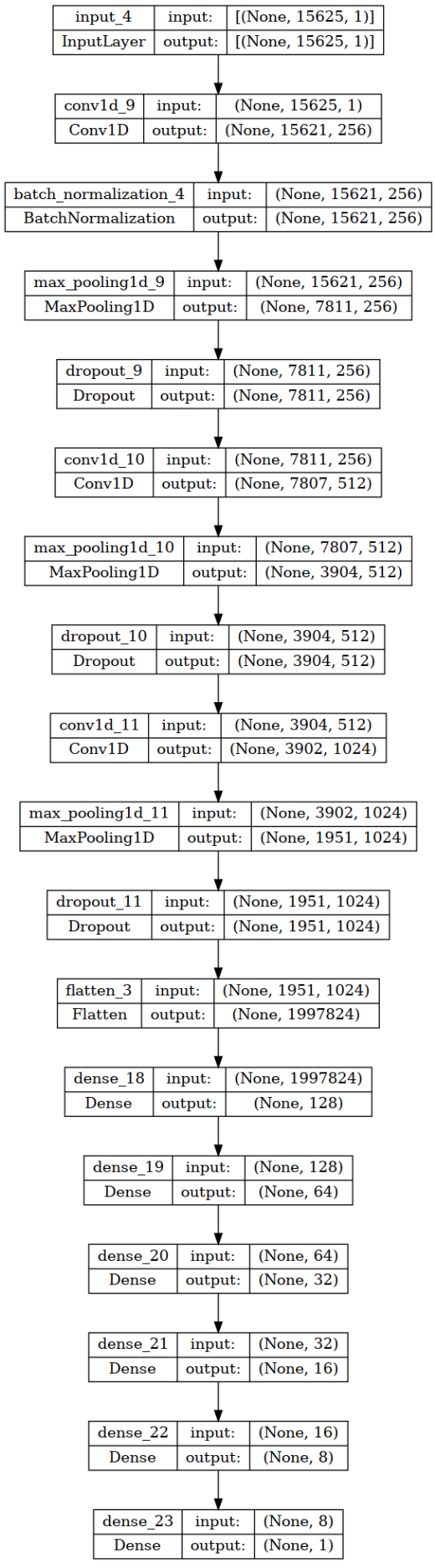

Figure 3.4: Hexamer Model Architecture

3.5.2 Performance Evaluation

The results from the Hexamer branch were :

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.3832695	0.83908	0.917847	0.7448275	648	58	812	222

3.6 Integration of Branches

The integration of the branches was driven by several key objectives. First and foremost, the aim was to improve the overall accuracy of translation initiation start position prediction by harnessing the collective power of multiple sources of information. By combining separate sequence patterns, secondary structure, conservation, and hexamer features, the integrated model sought to uncover and leverage the synergies between these components. Also, by integrating the several branches with one another, it can be observed which features each branch is able to extract from the initial input. This analysis helps to understand if all the branches are providing new and different inputs to the collective model, thereby enhancing the final prediction, or if some branches only add overhead to the model's predictions without offering distinct perspectives.

Additionally, the integration aimed to enhance precision and recall, ensuring a higher degree of correctness and sensitivity in identifying true translation initiation start positions. This objective was particularly crucial in reducing false positive and false negative predictions, which have significant implications for downstream analyses and understanding gene expression. Lastly, the integration sought to bolster the model's robustness, making it more resilient to noise, uncertainties, and variations in the input data. By utilizing the strengths of each branch and integrating them cohesively, the model aspired to improve prediction performance, advance the field of translation initiation start position prediction, and contribute to the understanding of gene expression and protein synthesis.

To integrate the branches effectively, a multi-branch fusion network approach was employed. The outputs from each branch were combined to create several prediction models that leverage the diverse information captured by the RNA Sequence Branch, Contrafold Branch, Conservation Branch, and Hexamer Branch, depending the integration. The fusion process involved aggregating the feature representations from each branch and feeding them into a shared set of fully connected layers. This approach facilitated the fusion of sequence pat-

terns, secondary structure information, conservation levels, and hexamer features, enabling the models to capture a comprehensive understanding of translation initiation start positions. By leveraging this integration, the models aimed to enhance prediction accuracy by leveraging the complementary nature of the branches and improving their ability to discern the critical features associated with translation initiation.

3.6.1 RNA Sequence + Contrafold Integration

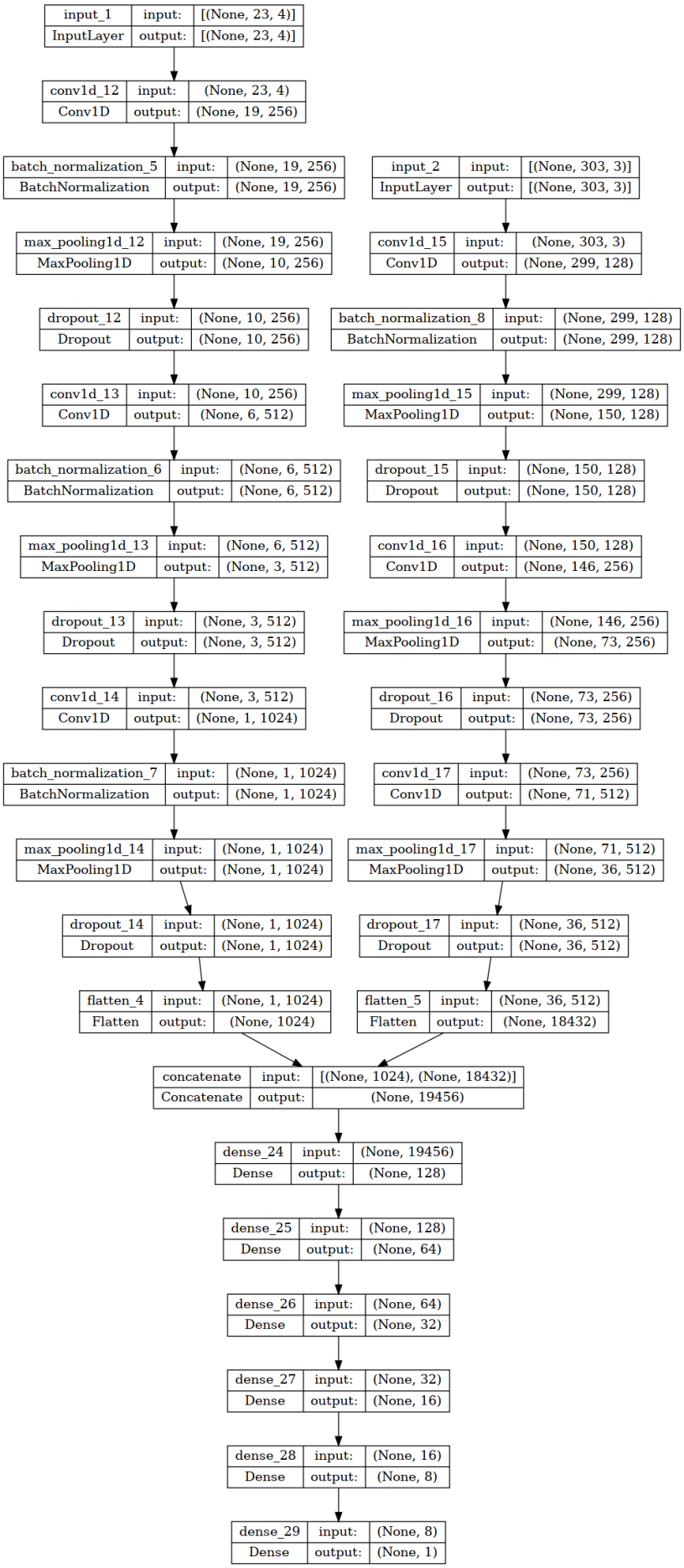

Figure 3.5: Sequence + Contrafold Model Architecture

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.392805	0.837931	0.87215	0.791954	689	101	769	181

The integration of the sequence and Contrafold branches aimed to leverage the additional information captured by the Contrafold branch, specifically the secondary structure characteristics, in combination with the sequence-based features. However, upon analyzing the results, it was observed that the Contrafold branch did not bring any new distinct features or substantial improvements in performance when combined with the sequence branch. The performance metrics and evaluation results revealed that the predictions made by the integrated model were comparable to those obtained solely using the sequence branch. This indicates that the Contrafold branch, despite its inclusion, did not provide any significant additional insights or contribute distinctive information for predicting translation initiation start positions. Therefore, the analysis suggests that the sequence branch alone is sufficient for capturing the relevant features and achieving accurate predictions in this context.

3.6.2 RNA Sequence + Conservation Integration

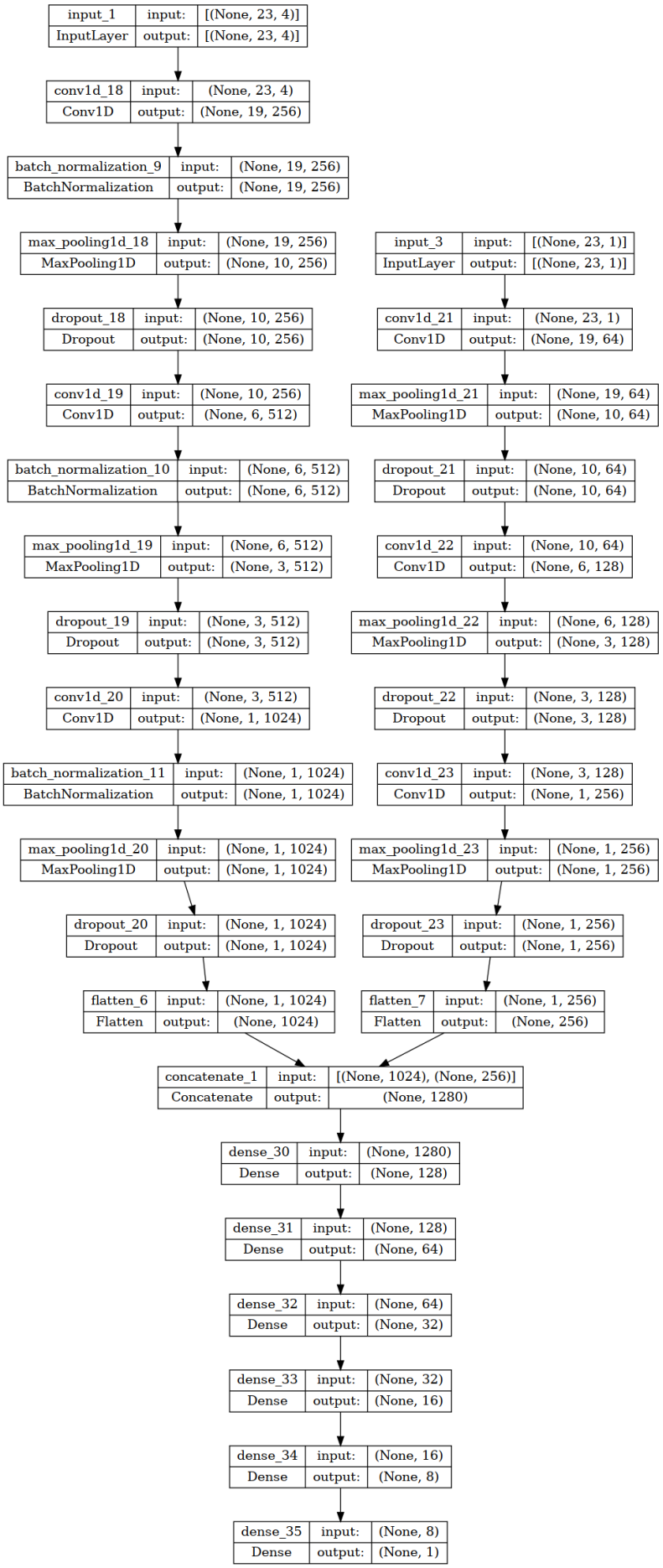

Figure 3.6: RNA Sequence + Conservation Model Architecture

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.4603286	0.7988505	0.823383	0.760919	662	142	728	208

The integration of the sequence and Conservation branches aimed to combine the sequence-based features with the conservation information extracted from the genomic context, with the expectation of improving the prediction accuracy of translation initiation start positions. However, upon analyzing the results, it was observed that the integration of the Conservation branch with the sequence branch did not yield the expected improvement and, in fact, resulted in slightly worse performance compared to using the sequence branch alone. The evaluation metrics and analysis revealed a marginal decrease in accuracy, precision, and recall, indicating that the incorporation of the conservation information did not contribute significantly to enhancing the prediction of translation initiation start positions. These findings suggest that, in this particular context, the Conservation branch did not provide additional valuable insights or complement the sequence-based features in capturing relevant patterns associated with translation initiation. As a result, it can be concluded that, for this specific task, the sequence branch alone remains the primary contributor to accurate predictions of translation initiation start positions.

3.6.3 Contrafold + Conservation Integration

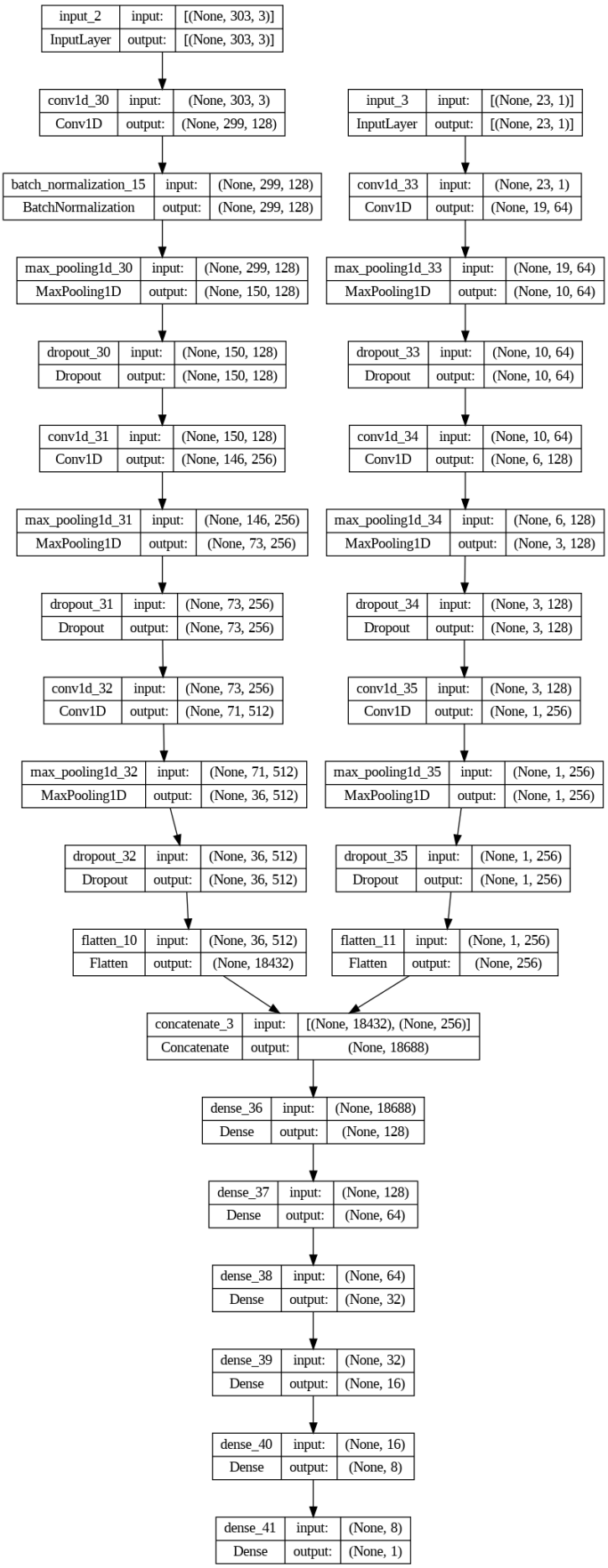

Figure 3.7: Contrafold + Conservation Model Architecture

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.3953962	0.8195402	0.8475	0.77931	678	122	748	192

The integration of the Contrafold and Conservation branches, achieved through concatenating their respective features, revealed intriguing insights into the prediction of translation initiation start positions. Notably, the combined integration demonstrated improved performance compared to the individual branches when considered in isolation. This enhancement in prediction accuracy, precision, and recall suggests a potential synergy between the secondary structure characteristics captured by the Contrafold branch and the evolutionary conservation patterns extracted by the Conservation branch. Interestingly, while the Contrafold and Conservation branches did not individually outperform the sequence branch, their concatenation still yielded almost similar results.

The improved performance resulting from the concatenated integration raises an intriguing possibility: that the Contrafold and Conservation branches when combined, contribute a complex feature that is not readily apparent in the sequence branch alone. The distinct characteristics captured by the Contrafold branch, pertaining to secondary structure, and the insights gleaned from the Conservation branch's focus on evolutionary conservation, seemingly complement each other in a manner that enhances the predictive power of the model. This suggests that the interplay between secondary structure and conservation features contributes to a nuanced and comprehensive understanding of translation initiation start positions that transcend the information provided by the sequence branch.

The observation of enhanced performance through the concatenation of the Contrafold and Conservation branches underscores the value of exploring diverse sources of information and leveraging their interactions. This finding not only underscores the complexity of the prediction task but also emphasizes the potential of integrative models in capturing intricate and multifaceted features crucial for the accurate prediction of translation initiation start positions.

3.6.4 RNA Sequence + Contrafold + Conservation Integration

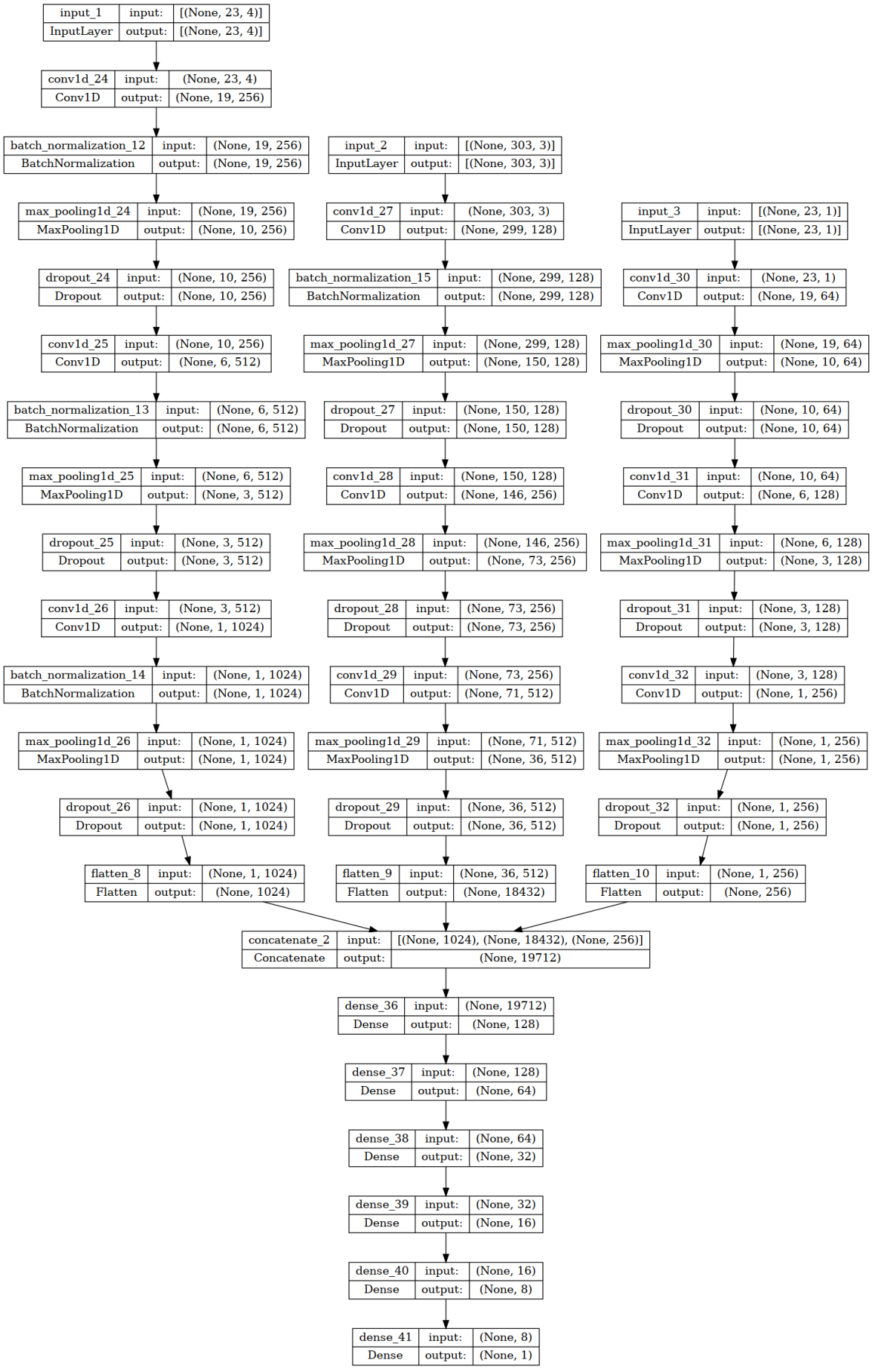

Figure 3.8: RNA Sequence + Contrafold + Conservation Model Architecture

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.2976811	0.882758	0.9012048	0.85977	748	82	788	122

The integration of the sequence, Contrafold, and Conservation branches proves to be the most effective configuration yet, outperforming both individual branches and previous dual-branch combinations. Notably, the combined integration of the Contrafold and Conservation branches with the sequence branch yields notably enhanced results. This integration seems to introduce fresh and complementary information, contributing to the model's improved precision in predicting translation initiation start positions. The synergy between secondary structure insights and evolutionary conservation patterns appears to unveil intricate data relationships, ultimately leading to a substantial performance enhancement.

3.6.5 RNA Sequence + Contrafold + Conservation + Hexamer Integration

Figure 3.9: RNA Sequence + Contrafold + Conservation + Hexamer Model Architecture

Loss	Accuracy	Precision	Recall	TP	FP	TN	FN
0.235511	0.905172	0.875399	0.9448275	822	117	753	48

The integration of the sequence, Contrafold, Conservation, and Hexamer branches marks the culmination of the exploration, yielding the highest accuracy at an impressive 90%. This

integrated model not only achieves the best accuracy but also boasts the lowest loss and the highest recall among all configurations tested. While it slightly lags in precision compared to the previous model, the holistic combination of these branches emerges as the clear leader. In essence, this comprehensive integration stands out as the most successful model, showcasing the potential of leveraging diverse features to achieve a robust and accurate prediction of translation initiation start positions.

Chapter 4

Experimental Results and Analysis

With the architectures of the predictive models now firmly established, this chapter turns its focus to empirical validation and analysis. A comprehensive examination of the models' performance across various dimensions will be presented, assessing their ability to accurately predict TIS positions. The evaluation metrics deployed provide a multi-faceted assessment of the models' effectiveness, from their ability to discern true positives and false positives to the intricate balance between precision and recall.

4.1 Evaluation Metrics

In the realm of classifier models for translation initiation start, the significance of evaluation metrics becomes paramount. These metrics act as a guiding compass, enabling us to navigate the intricate landscape of model performance. The nuanced nature of translation initiation start prediction requires a comprehensive evaluation that goes beyond mere accuracy. Metrics such as precision, and recall score delve into the model's ability to correctly identify true translation start sites while minimizing false positives. The confusion matrix, in particular, provides a clear breakdown of true negatives, true positives, false negatives, and false positives. By scrutinizing these metrics, we gain a holistic understanding of the model's strengths and areas for improvement. This comprehensive assessment empowers us to refine the model, addressing specific weaknesses and optimizing its predictive prowess. In a field where precision is pivotal, a thorough evaluation not only validates the model's capabilities but also elucidates its potential to enhance translation initiation site prediction.

Confusion Matrix

Model	Loss	Accuracy	Precision	Recall

Model	True Positives	False Positives	True Negative	False Negative

The evaluation begins, by presenting a detailed view of model performance through the confusion matrix. The confusion matrix provides a comprehensive breakdown of classification outcomes, revealing the interplay between true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) across different models.

Model Performance Metrics

In the presented confusion matrixes, the key performance metrics for each model have been summarized, including accuracy, precision, and recall. These metrics offer a clear indication of the model's ability to make accurate predictions and its sensitivity to identifying positive instances.

Loss: Assessment of the model's deviation from the ground truth. Lower loss values indicate better alignment between predicted and actual values.

Accuracy: The ratio of correctly predicted instances to the total instances, giving an overall performance measure.

Precision: The ratio of true positives to the sum of true positives and false positives, reflects the model's ability to avoid false alarms.

Recall: The ratio of true positives to the sum of true positives and false negatives, indicates the model's ability to capture all relevant positive instances.

The confusion matrix's components provide deeper insights into the classification process:

True Positives (TP): The number of correctly predicted positive instances.

False Positives (FP): The number of instances predicted as positive but are actually negative.

True Negatives (TN): The number of correctly predicted negative instances.

False Negatives (FN): The number of instances predicted as negative but are actually positive.

Significance in TIS Prediction

In the context of Translation Initiation Start (TIS) prediction, each metric holds special significance:

True Positives and Recall: A high recall signifies the model's capacity to capture actual TIS positions, minimizing the risk of missing important sites.

Precision: High precision ensures that predicted TIS positions are likely to be accurate, minimizing false positives that could lead to incorrect annotations.

These metrics are vital in evaluating the performance of the models' ensemble, helping gauge how effectively the system identifies TIS positions while maintaining a balance between precision and recall.

The True Positives vs. False Positives (TP vs. FP) plot, featuring a dynamic true positive threshold, introduces a distinct perspective into model classification performance. This particular metric serves a unique purpose in the evaluation, shedding light on how the model's behavior adapts as the classification threshold becomes increasingly conservative. In this plot, the transition of true positives occurs as the true positive threshold ranges from 0.7 to 1.0. This incremental change in the threshold allows for an exploration of the models' sensitivities to different decision criteria. The plot offers insights into how the models respond to increasingly stringent criteria for classifying instances as positive. As the threshold becomes more conservative and approaches 1.0, the models demonstrate a reluctance to classify instances as positive unless they are highly certain. This signifies a cautious approach, wherein the models' positive predictions only occur when they have a very high level of confidence. This conservatism can result in higher precision but may lead to a reduced recall, as the models become selective in labeling instances as positive. This dynamic threshold analysis

mirrors practical scenarios where decision thresholds must align with specific objectives. In the context of Translation Initiation Start (TIS) prediction, a highly conservative threshold can be employed when minimizing false positives is paramount. Researchers may require exceptionally strong evidence from the model before classifying an instance as positive.

The TP vs. FP plot with a dynamic threshold provides valuable insights into the models' adaptability and versatility, especially as the threshold becomes more conservative. It offers a clear depiction of how the models become increasingly cautious and selective in their positive predictions. Researchers can use this information to make informed decisions about the threshold that best suits their specific objectives, whether emphasizing precision, recall, or achieving a finely tuned balance.

The Precision vs. Sensitivity plot serves as a comprehensive metric for evaluating model performance, offering valuable insights into how the models navigate the intricate landscape of Translation Initiation Start (TIS) prediction. The plot's depiction of precision and sensitivity provides a dynamic view of how the models balance between correctly identifying true TIS positions and minimizing false positives. The Precision vs. Sensitivity plot quantifies model performance and visually illustrates the trade-offs and decision-making dynamics inherent in the models. By examining the curve's shape and trajectory, we gain insights into how the models prioritize precision over sensitivity and vice versa.

4.2 Baseline Models

4.2.1 Baseline Model Overview

In the realm of bio-informatics and genomics, traditional machine learning models have long played a pivotal role in various sequence analysis tasks. Among these, Support Vector Machines (SVM) and small feed-forward Multi-layer Perceptron Artificial Neural Networks (MLP ANN) stand out as notable examples. SVM, known for its ability to handle complex classification tasks through the use of well-defined decision boundaries, and MLP ANN, a classic neural network architecture with multiple layers of interconnected nodes, have been utilized in diverse biological sequence analysis scenarios. While these models have shown their mettle in several applications, it is essential to understand their principles and limitations when considering their application to specific tasks, such as predicting translation initiation start positions.

4.2.2 Datasets and Limitations

One significant challenge in comparing with the baseline machine learning models like SVM and MLP ANN lies in the differences between the datasets used for training and evaluation. SVM and MLP ANN models reported in the literature used datasets that were old and mostly have been constructed by various biology labs. Consequently, adapting these baseline models to a specific dataset can be non-trivial due to these variations. The model's performance may also depend heavily on the quality and representativeness of the dataset, making direct comparisons challenging when the datasets diverge significantly from one another.

4.2.3 Scalability and Interpretability

Traditional machine learning models, including SVM and MLP ANN, have both strengths and limitations. While they are generally interpretable, allowing researchers to glean insights into the decision-making process, they may face scalability issues when dealing with large-scale datasets or attempting to capture intricate, high-dimensional patterns. As a result, the applicability of these models can vary depending on the task's complexity and the availability of extensive labeled data. Deep learning models, on the other hand, have demonstrated a remarkable capacity to automatically extract meaningful features from raw data, making them suitable for tasks where the underlying patterns are highly complex or poorly understood.

Applicability and Conclusion: In conclusion, while SVM and MLP ANN have been valuable tools in bioinformatics and genomics, their applicability and performance should be considered in light of the unique challenges posed by the task at hand. The fundamental differences in the learning mechanisms between traditional machine learning models and deep learning models, such as the one developed in this study, underscore the need to tailor the choice of model to the problem's specific requirements. While traditional machine learning models are interpretable and have their merits, the automated feature learning capabilities of deep learning models offer promise in tackling complex, data-rich biological sequence analysis tasks, such as predicting translation initiation start positions.

4.3 State-of-the-Art Models

The accurate prediction of translation initiation start sites (TIS) from genomic sequences represents a crucial endeavor in genomics and bioinformatics. Understanding the mecha-

nisms of gene expression and regulation hinges upon the ability to precisely identify these initiation sites. In the past years, extensive research has been dedicated to this subject, driven by its significance in unraveling gene function and regulation. As a result, a plethora of computational methods have emerged, ranging from traditional machine learning approaches to cutting-edge deep learning models. In this section, three state-of-the-art models will be examined and elucidated as to why they are particularly well-suited for addressing the TIS prediction problem.

In recent literature focusing on TIS prediction, a prevailing approach is the utilization of a combined architecture that incorporates both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Models such as GCR-Net, NeuroTIS, DeepTIS, and TITER have all embraced this approach, sharing a common framework aimed at enhancing TIS prediction.

While these models exhibit subtle variations in their design details, they all share the central idea of harnessing the capabilities of CNNs and RNNs to effectively capture patterns in genomic sequences. These patterns are critical for accurately predicting translation initiation sites.

One noteworthy distinction among these models is the sequence in which they employ CNN and RNN layers. Some models, like GCR-Net and DeepTIS, initiate with CNN layers, capturing spatial features before handing off the processed data to RNN layers for temporal dependencies. In contrast, models such as NeuroTIS and TITER prioritize temporal dependency modeling with RNN layers before incorporating spatial information using CNN layers. These architectural nuances, though minor, contribute to their respective approaches. Additionally, while some models incorporate k-mer frequencies taken from the data as the only feature, others rely solely on genomic sequence data.

Despite the presence of minor variations among these models, it's notable that their collective efforts essentially converge on a shared objective. While they may appear distinct in certain design details, their overarching approach remains largely uniform. As such, it's important to acknowledge that these models, due to their substantial similarities, do not substantially expand the knowledge base in the field of TIS prediction. Instead, they reiterate existing methodologies and insights. Consequently, their contribution to the scientific field's ongoing efforts in TIS prediction is relatively limited, as they primarily reinforce the already established notion that a CNN-RNN hybrid architecture can be efficacious in addressing the

unique challenges of predicting translation initiation sites in genomic sequences.

4.4 Discussion of Results

Model	Loss	Accuracy	Precision	Recall
RNA Sequence	0.3862122	0.83333	0.846062	0.814942
Contrafold	0.536843	0.73103	0.68075	0.870114
Conservation	0.44620147	0.776436	0.904201	0.618390
Hexamer	0.3832695	0.83908	0.917847	0.7448275
RNA Sequence + Contrafold	0.392805	0.837931	0.87215	0.791954
RNA Sequence + Conservation	0.4603286	0.7988505	0.823383	0.760919
Contrafold + Conservation	0.3953962	0.8195402	0.8475	0.77931
RNA + Contrafold + Conservation	0.2976811	0.882758	0.9012048	0.85977
Total	0.235511	0.905172	0.875399	0.9448275

Model	TP	FP	TN	FN
RNA Sequence	709	129	741	161
Contrafold	757	355	515	113
Conservation	538	57	813	332
Hexamer	648	58	812	222
RNA Sequence + Contrafold	689	101	769	181
RNA Sequence + Conservation	662	142	728	208
Contrafold + Conservation	678	122	748	192
RNA + Contrafold + Conservation	748	82	788	122
Total	822	117	753	48

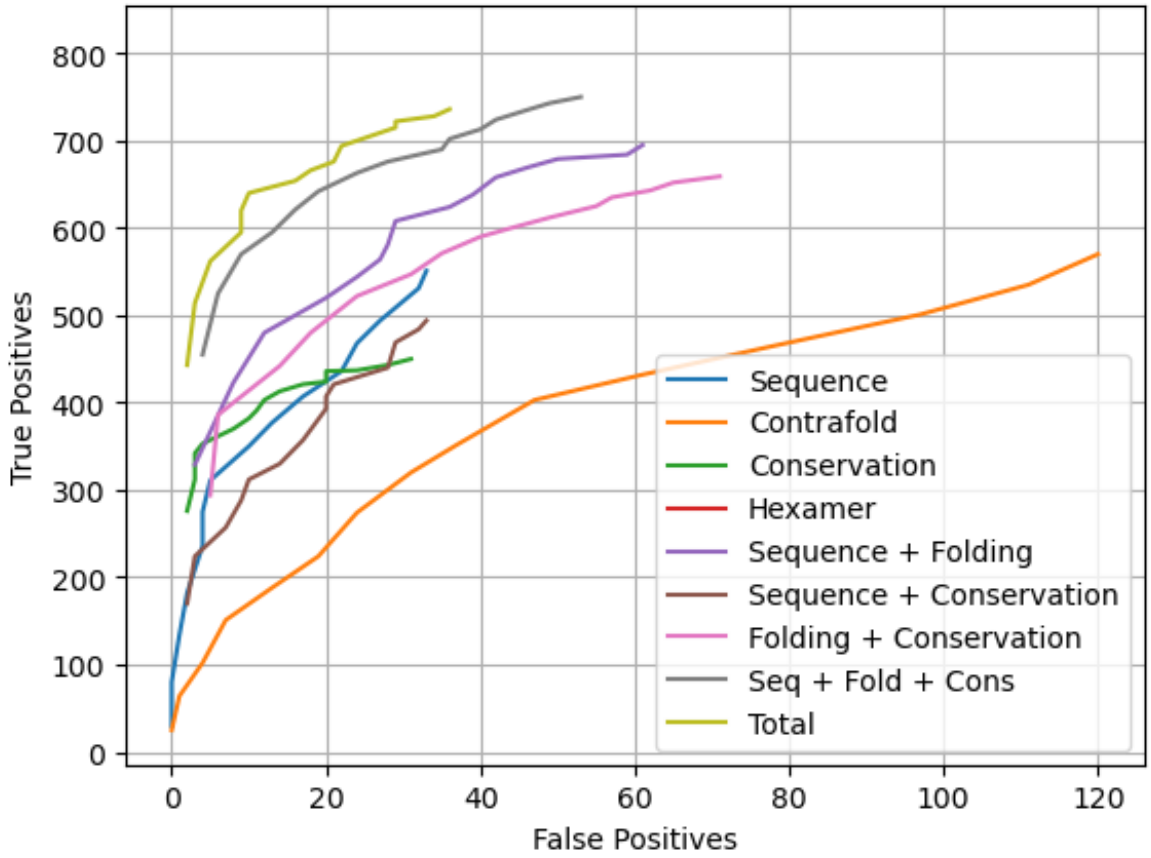

Figure 4.1: True Positives VS False Positives (for each Network)

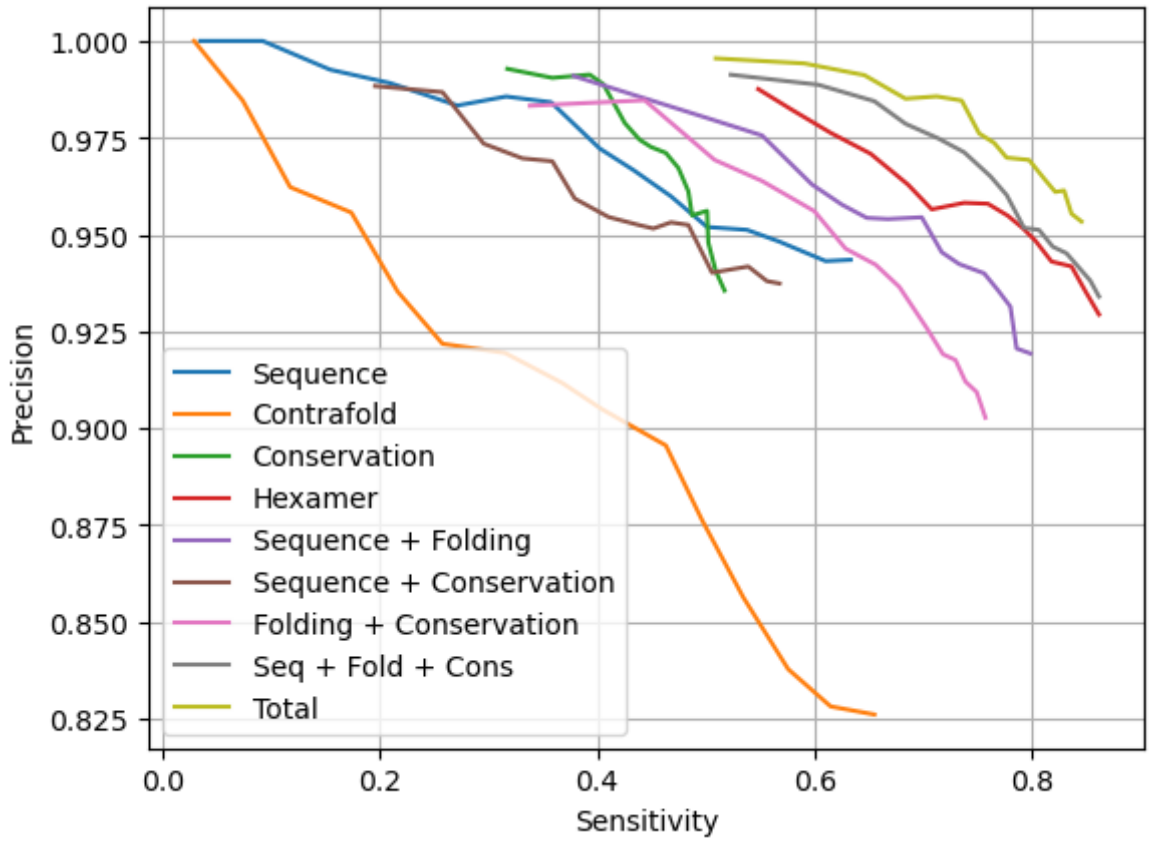

Figure 4.2: Precision VS Recall (for each Network)

Analysis of the confusion matrix reveals intriguing insights into the impact of various features on model performance. Notably, the RNA sequence and hexamer features demonstrate a capacity to yield the most favorable results individually. However, an intriguing observation emerges when combining the Contrafold and conservation features. The fusion of these branches produces results that surpass individual contributions, indicating their ability to capture distinct patterns within the data, thereby enhancing the model's predictive capabilities. This observation underscores the complexity of TIS prediction, suggesting the existence of nuanced intricacies yet to be fully understood.

Furthermore, our results demonstrate that the model is adept at learning diverse patterns from distinct feature sets, including the RNA Sequence, the combined Contrafold and Conservation features, and the hexamer feature.

The true positives vs. false positives plot, where we manipulate the threshold for positive predictions (setting it above 0.5), and the precision vs. recall plot both reaffirm the idea that incorporating multiple features leads to improved model performance. Additionally, a noteworthy trend emerges in the true positives vs. false positives plot: as models become more

conservative, their performance metrics notably improve. This alignment between improved metrics and a conservative approach is mirrored in the precision vs. sensitivity plot, reinforcing the idea that leveraging multiple features enhances predictive accuracy and robustness.

It is worth noting that the proposed model demonstrates superior performance compared to the baseline models, namely the SVM and the small dense MLP. This improved performance can be attributed to several factors. Firstly, the proposed model benefits from a deeper architecture, which allows it to capture more complex patterns in the data. Secondly, the use of convolutional neural networks (CNNs) proves to be advantageous in this context, as they excel in extracting features from sequential data, aligning well with the nature of the TIS prediction problem. Lastly, our model leverages an extensive set of features derived from the data, providing it with richer information for decision-making. Collectively, these advantages contribute to the notable success of our proposed model in outperforming the baseline models.

Finally, the evaluation of other existing models on our dataset or the application of our model on datasets tailored for those methods was challenged by the absence of readily available models or final datasets for comparison. This limitation impeded the ability to conduct a fair evaluation by implementing the proposed model with the prescribed architecture and testing it against these methods.

Chapter 5

Conclusion and Future Work

5.1 Summary of Contributions

The contributions of this thesis can be divided into two, the dataset and the proposed model.

5.1.1 Dataset

The dataset created for this research represents a fundamental cornerstone of this thesis's contributions. It encompasses a carefully curated collection of RNA sequences, meticulously structured to facilitate the accurate prediction of Translation Initiation Start (TIS) positions. The significance of a well-annotated dataset in genomics research cannot be overstated. It underpins the reliability of any predictive model or analysis, as the quality of the data directly impacts the validity of the findings. In the context of predicting TIS positions, precise annotations are paramount to accurately discerning the initiation sites in RNA sequences. A well-annotated dataset ensures that the training and evaluation of machine learning models are based on ground truth information, reducing the likelihood of spurious results and false discoveries. Additionally, it facilitates the reproducibility of research efforts, enabling other scientists to build upon the dataset and contribute to the advancement of knowledge in the field. This dataset's meticulous curation and annotation set a high standard for future research in the domain of TIS prediction. Ensuring the accessibility of the dataset is a core tenet of this research's mission. By making the dataset available to the scientific community, it becomes a valuable resource for researchers and data scientists working in genomics, bioinformatics, and related fields.

5.1.2 Proposed Model

The cornerstone of this thesis lies in the development of a robust and accurate deep learning model for the prediction of Translation Initiation Start (TIS) positions in the human genome. This model represents a significant contribution to the field of genomics and bioinformatics. Its high accuracy, averaging around 90% with a narrow margin of error, demonstrates its practical viability and effectiveness. Furthermore, the proposed model is openly available for download, providing researchers worldwide with easy access to its powerful predictive capabilities. Finally, in summary, the proposed model not only represents a significant advancement in TIS prediction but also stands as a valuable resource that can catalyze further discoveries in genomics. Its accessibility ensures that its potential benefits can be realized by researchers worldwide, ultimately driving innovation and progress in the ever-evolving landscape of genomic research.

5.2 Limitations of the Proposed Model

One notable limitation of the model is its increased RAM consumption due to the implementation of k-mers, which can make it resource-intensive for some computational setups. Additionally, while the model exhibits commendable accuracy, it is important to acknowledge that TIS prediction remains a challenging task, and despite its strong metrics, it has not yet reached the coveted 95%+ accuracy threshold, leaving room for further refinement and exploration of advanced techniques in the future.

5.3 Future Directions and Enhancements

Looking ahead, several promising avenues are seen for further enhancement of both the model's architecture and the dataset used in this research.

5.3.1 Model Enhancements

In terms of the model itself, one compelling direction to explore is the integration of attention mechanisms, like transformer layers. Attention mechanisms have demonstrated remarkable potential in various natural language processing and sequence-related tasks. Integrating

attention into the model's architecture could provide it with the ability to focus on relevant features within the genomic data, potentially improving its predictive power.

Another key area for improvement is the development of explainability tools for the model. An interpretable model can help researchers understand how the features are being used and leveraged by the model. This can be particularly valuable when attempting to gain a deeper understanding of the intricate relationships between RNA secondary structure, evolutionary conservation, and TIS prediction. Exploring explainable AI methods can provide a clearer picture of which features are most influential in the decision-making process of the model.

5.3.2 Future Directions

The application of the proposed model to the entire human genome represents a compelling prospect for future research endeavors. Extending the model's reach to the entirety of the human genome offers the opportunity to not only verify and validate known Translation Initiation Start (TIS) positions but also to uncover previously unexplored regions of interest. This comprehensive exploration can serve a dual purpose: firstly, to confirm the accuracy of the model's predictions in known TIS sites, thereby bolstering its practical utility in the identification of experimental proteins; and secondly, to identify potential novel TIS candidates in hitherto uncharted genomic territories. These uncharted regions hold the potential to unravel hidden aspects of gene expression regulation, presenting researchers with fresh avenues for investigation and experimentation. By venturing beyond the confines of known TIS sites, this expansion of scope offers a rich opportunity to advance our understanding of translation initiation and its intricate role in shaping the proteome.

In closing, the journey of refining the model and exploring the uncharted territories of the human genome holds great promise for the field of genomics. These future directions represent a pivotal step forward in our quest to decipher the complex mechanisms governing translation initiation and its impact on the proteome.

Bibliography

[1] Artemis Hatzigeorgiou and Molly Megraw. *Computational Analysis of Human DNA Sequences: An Application of Artificial Neural Networks*, pages 169–180. Springer US, Boston, MA, 2006.

[2] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic dna11edited by f. e. cohen. *Journal of Molecular Biology*, 268(1):78–94, 1997.

[3] Christian Derst, Martin Reczko, and Artemis Hatzigeorgiou. Prediction of human translational initiation sites using a multiple neural network approach, 2000.

[4] A Zien, G Rätsch, S Mika, B Schölkopf, and T Lengauer. Engineering support vector machine kernels that recognize translation initiation sites, 2000.

[5] S Lander et al. Initial sequencing and analysis of the human genome international human genome sequencing consortium* the sanger centre: Beijing genomics institute/human genome center, 2001.

[6] Artemis G Hatzigeorgiou. Translation initiation start prediction in human cdnas with high accuracy, 2002.

[7] Donna Karolchik, Angela S. Hinricks, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The ucsc table browser data retrieval tool. *Nucleic Acids Research*, 32, 1 2004.

[8] Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. volume 22. Oxford University Press, 7 2006.

[9] Axel Bernal, Koby Crammer, Artemis Hatzigeorgiou, and Fernando Pereira. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Computational Biology*, 3:0488–0497, 2007.

[10] Aaron R. Quinlan and Ira M. Hall. Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 1 2010.

[11] Ian Dunham et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 9 2012.

[12] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 04 2012.

[13] Alan G. Hinnebusch, Ivaylo P. Ivanov, and Nahum Sonenberg. Translational control by 5′-untranslated regions of eukaryotic mrnas, 6 2016.

[14] Sai Zhang, Hailin Hu, Tao Jiang, Lei Zhang, and Jianyang Zeng. Titer: Predicting translation initiation sites by deep learning. volume 33, pages i234–i242. Oxford University Press, 7 2017.

[15] Georgios K. Georgakilas, Andrea Grioni, Konstantinos G. Liakos, Eliska Chalupova, Fotis C. Plessas, and Panagiotis Alexiou. Multi-branch convolutional neural network for identification of small non-coding rna genomic loci. *Scientific Reports*, 10, 12 2020.

[16] Chao Wei, Junying Zhang, Xiguo Yuan, Zongzhen He, Guojun Liu, and Jinhui Wu. Neurotis: Enhancing the prediction of translation initiation sites in mrna sequences via a hybrid dependency network and deep learning framework. *Knowledge-Based Systems*, 212, 1 2021.

[17] Dimitris Grigoriadis, Nikos Perdikopanis, Georgios K. Georgakilas, and Artemis G. Hatzigeorgiou. Deeptss: multi-branch convolutional neural network for transcription start site identification from cage data. *BMC Bioinformatics*, 23, 12 2022.

[18] Weihua Li, Yanbu Guo, Bingyi Wang, and Bei Yang. Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction. *Pattern Recognition*, 136, 4 2023.

[19] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.