UNSUPERVISED ALGORITHMS FOR AUTOMATED GENE PREDICTION IN NOVEL EUKARYOTIC GENOMES

A Dissertation
Presented to
The Academic Faculty

By

Tomáš Brůna

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics
School of Biological Sciences

Georgia Institute of Technology

August 2022

UNSUPERVISED ALGORITHMS FOR AUTOMATED GENE PREDICTION IN NOVEL EUKARYOTIC GENOMES

Thesis committee:

Dr. Mark Borodovsky, Advisor
School of Computational Science and Engineering and Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Xiuwei Zhang
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Kostas T. Konstantinidis
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Jung H. Choi
School of Biological Sciences
Georgia Institute of Technology

Date approved: July 18, 2022

To my parents, Tomáš and Jaromíra,

my brother, Lukáš,

and my wife, Parastoo

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my advisor, Dr. Mark Borodovsky, for his guidance, support, and for providing me with the opportunity to collaborate with many researchers both within and outside of Georgia Tech. Thanks to Alex Lomsadze for his encouragement and countless pieces of useful advice. I thank Karl Gemayel and Aaron Pfennig for their friendship and valuable scientific discussions. I am also grateful to my committee members, Dr. King Jordan, Dr. Jung Choi, Dr. Xiuwei Zhang, and Dr. Kostas Konstantinidis, for their insights and for reviewing this thesis. Special thanks are due to the team from the University of Greifswald, Katharina Hoff, Mario Stanke, and Lars Gabriel, for a fruitful collaboration that resulted in much of the work presented in this dissertation. I gratefully acknowledge the financial support provided by the National Institutes of Health grant GM128145 and the Naumann-Etienne Foundation Fellowship. Many thanks to Lisa Redding; I cannot imagine having a more helpful academic coordinator. Thanks should also go to my undergraduate advisor, Tomáš Bartoň, whose vast knowledge and work ethic inspired me to pursue a graduate degree. I would like to thank my dear friends from Unit 190, Oğulcan Canbek, Patrick Friedrich, and Elio Challita, for offering excellent feedback on my research presentations. I also wish to thank my brother, Lukáš, and my parents, Tomáš and Jaromíra, for their love, continuous support, and push towards education. Last but not least, I cannot begin to express my appreciation to my wife, Parastoo, who is always there for me and who supported me through every step of my doctoral studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AEE Alignment of an Entire Exon

BAQ Border Alignment Quality

F1 harmonic mean

FN false negatives

FP false positives

GHMM generalized hidden Markov model

HC high-confidence

HSSM hidden semi-Markov model

IBA Intron Borders Alignment

IMC Intron Mapping Coverage

mRNA messenger RNA

NGS next generation sequencing

ORF open reading frame

pre-mRNA precursor messenger RNA

RNA-Seq short-read RNA sequencing

SMC Site Mapping Coverage

Sn sensitivity

Sp specificity

TIS translation initiation signal

TP true positives

SUMMARY

Gene prediction, the identification of the location and structure of protein-coding genes in genomic sequences, is one of the first and most important steps in the analysis of assembled genomes. The exponential growth of sequenced eukaryotic genomes necessitates fully automated computational gene prediction methods. Due to the complexity and diversity of eukaryotic genomes, the task of accurate automatic eukaryotic gene prediction remains an open challenge. This work presents three novel gene prediction algorithms that address specific aspects of this challenge and thus improve over existing gene prediction methods.

The first part of this thesis describes GeneMark-EP+, an unsupervised gene prediction algorithm that uses homologous cross-species proteins to guide its model training and gene prediction steps. In contrast to existing homology-based gene finders, which can only extract information from proteins of closely related species, GeneMark-EP+ is designed to utilize proteins of any evolutionary distance, including remote homologs. Consequently, GeneMark-EP+ can fully exploit the information contained in large and ever-growing protein databases that are, unlike transcriptomic data, always readily available prior to a genome annotation project start. GeneMark-EP+ is shown to significantly improve over previous GeneMark versions, including ones integrating transcriptomic data.

In the second part, BRAKER2 is presented—a fully automated protein homology-based gene prediction pipeline that integrates GeneMark-EP+ with AUGUSTUS, an accurate gene finder that requires supervised training. By combining complementary strengths of these two gene prediction tools, BRAKER2 achieves state-of-the-art gene prediction accuracy in a fully unsupervised manner. The high gene prediction accuracy of BRAKER2 is demonstrated in tests on a wide range of plant and animal genomes. Further, it is shown that BRAKER2 compares favorably with MAKER2, one of the most popular gene prediction pipelines.

Finally, this thesis describes GeneMark-ETP+, a self-training gene prediction algo-

rithm that simultaneously utilizes diverse information streams—genomic, transcriptomic, and protein homology—throughout all stages of its model training and gene prediction. This evidence integration is achieved by, among other things, creating a novel method for simultaneous gene prediction in transcripts and genomic DNA. Notably, GeneMark-ETP+ builds upon the previous work of this thesis: its training is fully unsupervised and proteins of any evolutionary distance are utilized. The integrative approach of GeneMark-ETP+ is demonstrated to reach better prediction accuracy compared with competing tools combining *ab initio*-, protein homology-, and transcriptome-based predictions.

The research presented in this thesis contributed to the following publications:

Tomáš Brůna, Alexandre Lomsadze, Mark Borodovsky. "GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins". *Nucleic Acids Research Genomics and Bioinformatics* (2020), doi: https://doi.org/10.1093/nargab/lqaa026

Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky. "BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database". *Nucleic Acids Research Genomics and Bioinformatics* (2021), doi: https://doi.org/10.1093/nargab/lqaa108

Tomáš Brůna, Alexandre Lomsadze, Mark Borodovsky. "GeneMark-ETP+: automatic annotation of eukaryotic genomes in consistence with omics data". In preparation

Waleed MM El-Sayed, Alli L Gombolay, Penghao Xu, Taehwan Yang, Youngkyu Jeon, Sathya Balachander, Gary Newnam, Sijia Tao, Nicole E Bowen, Tomáš Brůna, Mark Borodovsky, Raymond F Schinazi, Baek Kim, Yongsheng Chen, Francesca Storici. "Disproportionate presence of adenosine in mitochondrial and chloroplast DNA of Chlamydomonas reinhardtii". *iScience* (2021), doi: https://doi.org/10.1016/j.isci.2020.102005

Lars Gabriel, Katharina J Hoff, Tomáš Brůna, Mark Borodovsky, Mario Stanke. "TSEBRA: transcript selector for BRAKER". *BMC Bioinformatics* (2021), doi: https://doi.org/10.1186/s12859-021-04482-0

Tomáš Brůna, Rishi Aryal, Olga Dudchenko, Daniel James Sargent, Daniel Mead, Matteo Buti, Andrea Cavallini, Timo Hytönen, Javier Andrés, Melanie Pham, David Weisz, Flavia Mascagni, Gabriele Usai, Lucia Natali, Nahla Bassil, Gina E Fernandez, Alexandre Lomsadze, Mitchell Armour, Bode Adebowale Olukolu, Thomas J Poorten, Caitlin Britton, Jahn Davik, Hamid Ashrafi, Erez Lieberman Aiden, Mark Borodovsky, Margaret Leigh Worthington. "A chromosome-length genome assembly and annotation of blackberry (Rubus argutus, cv. 'Hillquist')". *bioRxiv* (2022), doi: https://doi.org/10.1101/2022.04.28.489789

CHAPTER 1
INTRODUCTION

Gene prediction, the identification of the location and structure of protein-coding genes in genomic sequences, is one of the first and most important steps in the analysis of assembled genomes [1]. The significance of accurate gene prediction cannot be overstated because inaccurate predictions can hamper all downstream genomic analyses [2]; e.g., the functional investigation of an organism's genes, or even the analysis of related species [3].

In the past, accurate gene prediction was often achieved through manual curation and experimental validation of predicted gene models. Such a time- and resource-intensive approach, suitable for the then limited amount of sequenced genomic sequences, is not scalable. The exponential growth of eukaryotic assemblies (Figure 1.1), driven by constantly improving next-generation sequencing technologies, necessitates fully automated accurate computational gene prediction procedures [4]. While such procedures have already been successfully deployed for the computational annotation of prokaryotic genomes [7], the task of automatic gene prediction in eukaryotes remains an open problem [8, 9].

This thesis describes new computational methods addressing many of the open challenges in automatic eukaryotic gene prediction. Specifically, this thesis is organized as follows. Chapter 2 provides a background of gene prediction and defines the specific open gene prediction problems addressed by this thesis.

Chapter 3 describes GeneMark-EP+, a novel unsupervised gene prediction algorithm that uses homologous proteins to guide its model training and gene prediction steps. In contrast to existing homology-based gene finders, which are designed to extract information from proteins of closely related species, GeneMark-EP+ utilizes proteins of any evolutionary distance (including remote homologs) to better its predictions. The chapter shows that GeneMark-EP+ significantly improves over previous GeneMark versions, including

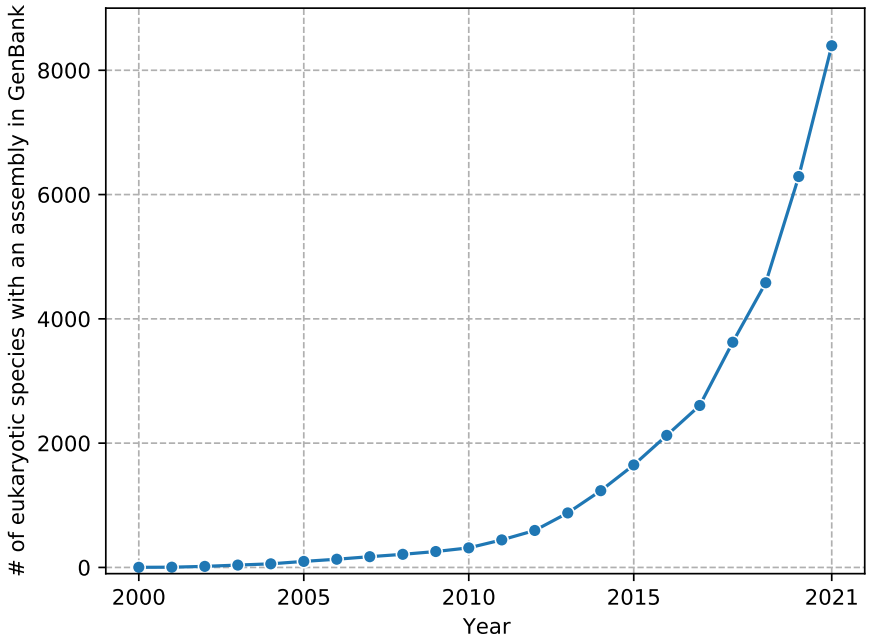

Figure 1.1: The growth of sequenced eukaryotic genomes. The data were collected from Genbank's genome table [5, 6].

ones integrating transcriptomic data. We also demonstrate that this improvement occurs even when proteins originating from evolutionarily remote species are used as input to GeneMark-EP+.

Chapter 4 describes BRAKER2, a fully automated protein homology-based gene prediction pipeline that integrates GeneMark-EP+ and other algorithms introduced in Chapter 3 with AUGUSTUS. By combining complementary strengths of multiple gene prediction tools, BRAKER2 achieves state-of-the-art gene prediction accuracy in a fully unsupervised manner. The chapter demonstrates this claim by evaluating BRAKER2's accuracy on a wide range of test genomes and comparing BRAKER2 with competing gene prediction pipelines.

Chapter 5 describes GeneMark-ETP+, a gene prediction algorithm that utilizes diverse information streams—genomic, transcriptomic, and protein homology—throughout all stages of its model training and gene prediction. The chapter demonstrates that the

integrative approach of GeneMark-ETP+ results in better prediction accuracy when compared to competing tools that combine multiple independent *ab initio*, homology-, and transcriptome-based predictions. Notably, the method presented in this chapter builds upon the work described in Chapter 3 and Chapter 4: its training is fully unsupervised, and the protein homology evidence integration utilizes proteins of any evolutionary distance.

Finally, Chapter 6 concludes this thesis by summarizing its contributions to the gene prediction field. Furthermore, the chapter lists several remaining unresolved challenges and suggests what future work could be done to solve them.

CHAPTER 2
BACKGROUND

This chapter provides the necessary background for Chapters 3 to 5. First, we define the task of eukaryotic gene prediction. Next, we describe the common sources of information used by gene prediction algorithms, along with examples of algorithms utilizing each given source. Subsequently, we describe open gene prediction problems addressed by this thesis, and finally, we define metrics used to assess gene prediction accuracy.

2.1 Gene prediction

Structural protein-coding gene prediction, also called structural genome annotation, is the identification of the location and structure of protein-coding genes in a genomic sequence. This is a distinct process from functional gene prediction/annotation which assigns a biological function to the predicted genes. For simplicity, we refer to "structural protein-coding gene prediction in eukaryotic genomes" as *gene prediction* in this thesis. The rest of this section describes the eukaryotic gene structure and uses the description to provide a more precise gene prediction definition.

2.1.1 Eukaryotic gene structure

Genes specify the structure of proteins, the building blocks of life. In eukaryotes, the biological process of expressing a protein amino acid sequence from a gene stored in DNA occurs in three major steps (Figure 2.1). First, the entire DNA sequence of a gene is *transcribed* to a *precursor messenger RNA (pre-mRNA)* transcript. The DNA gene and the corresponding pre-mRNA transcript are composed of two types of sequences, *exons* and *introns*. In the second step, the introns are excised from the pre-mRNA in a process called *splicing*. After the removal of introns, the exons are joined together to create a *messenger*

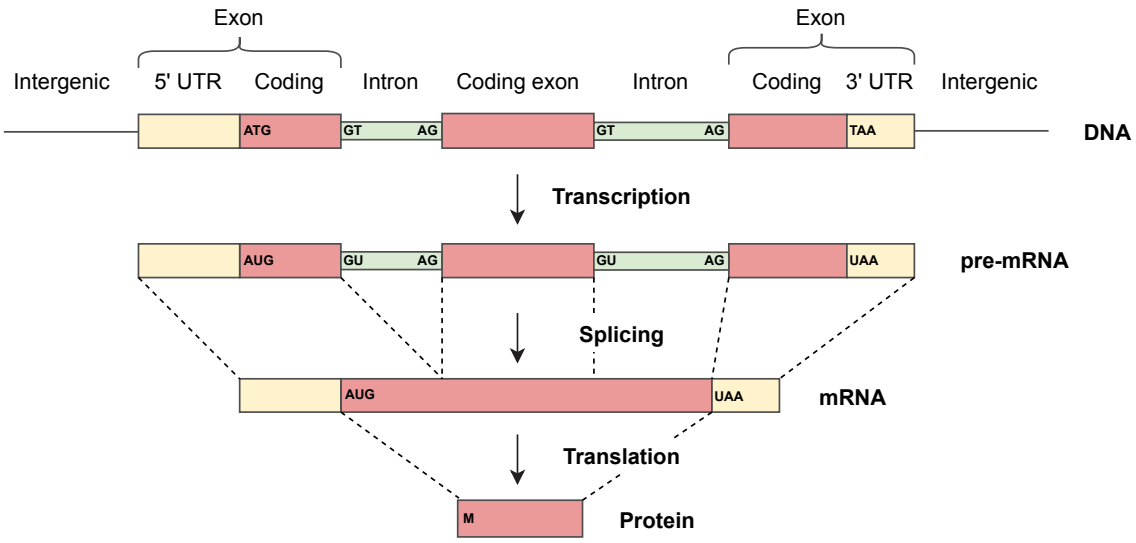

Figure 2.1: Illustration of the eukaryotic gene structure and the gene expression process.

RNA (mRNA) transcript. Finally, an internal part of the mRNA is *translated* to a protein. The translation occurs in triplets of nucleotides called *codons*; each codon codes for one amino acid or a stop signal. Typically, the translation begins with an AUG start codon and stops when it reaches one of the three stop codons (UAA, UAG, UGA).

2.1.2 Definition of gene prediction

With the description of a eukaryotic gene structure, we can define gene prediction more precisely. The task is to predict the genomic locations of all protein-coding genes delineated by the genomic positions of translation starts and stops in a DNA sequence. Due to the possible presence of introns, the precise locations of all exon-intron boundaries, called *splice sites*, must also be predicted. These coordinates fully define the resulting protein molecule.

Apart from the presence of introns, eukaryotic gene prediction is further complicated by alternative isoforms. A single gene can code for multiple alternative protein products, created through alternative processing of pre-mRNA. This thesis describes methods both with and without alternative isoform prediction capabilities.

2.2 Sources of information for gene prediction

Gene prediction algorithms generally utilize data from one or more of the following sources: (i) intrinsic statistical patterns of the genomic sequence itself, (ii) transcriptomic evidence from RNA sequencing, and (iii) protein homology. Notably, these information sources originate from the various stages of the gene expression process depicted in Figure 2.1.

2.2.1 Intrinsic evidence

Algorithms predicting genes solely from the intrinsic characteristics of the target genome are referred to as *ab initio* methods. *Ab initio* algorithms model and detect genomic signals used by the cell to express genes. These signals include, e.g., translation start patterns, splice site motifs, or branch points. Furthermore, *ab initio* algorithms model genome-specific characteristics such as the length distribution of exons and introns or the nucleotide composition of coding and non-coding regions. Most modern *ab initio* gene finders use a generalized hidden Markov model (GHMM) [10] (also referred to as HMM with duration [11] or hidden semi-Markov model (HSSM)) as their underlying statistical model. Examples of popular GHMM-based *ab initio* gene finders are AUGUSTUS [12], SNAP [13], GENSCAN [14], Fgenesh [15], or GeneMark-ES [16, 17].

The biggest strength of *ab initio* methods lies in their ability to predict genes based on the DNA sequence alone, in the absence of other evidence. *Ab initio* gene finders thus play a crucial role in the discovery of novel genes. Their weakness lies in the fact that the statistical signals of many genes are weak and hard to distinguish from noise [18]. As a result, the accuracy of purely *ab initio* gene finders is far from perfect, especially in large eukaryotic genomes [19–22]. For this reason, gene prediction tools often combine the *ab initio* component with other evidence which is extrinsic to the genomic DNA. Another challenge in utilizing the intrinsic evidence is caused by significant inter-species variance in statistical genomic patterns [23–26]. Accounting for this variation poses a problem which

is discussed in Section 2.3.1.

2.2.2 Transcriptomic evidence

The general idea behind transcriptome-based gene prediction algorithms is to directly read and map the mRNA transcript (see Figure 2.1) to the genomic DNA from which it was transcribed. The mRNA reading is most commonly done by short-read RNA sequencing (RNA-Seq) [27]. Because of the short length of RNA fragments delivered by RNA-Seq (typically 50–150 bp [28]), full-length transcripts need to be computationally reconstructed from the individual reads. This can be done by *de novo* assembling RNA reads with tools such as Trinity [29] or Oases [30]. However, more commonly in gene prediction, the individual short reads are first splice-aligned to the genome (e.g., by STAR [31] or HISAT2 [32]) and then assembled into full transcripts with tools such as StringTie [33, 34], PsiCLASS [35], or Cufflinks [36].

As an alternative to short-read RNA-Seq, the emerging long read sequencing technologies [37, 38] now enable the sequencing of full-length mRNA transcripts. Consequently, there is no need to computationally assemble such transcripts and they can be directly mapped to DNA, using tools such as GMAP [39] or Minimap2 [40]. However, compared to short-read RNA-Seq, long-read technology is more costly and exhibits higher sequencing error rates [41, 42].

The mapping of a full mRNA transcript to DNA (obtained by either of the methods described above) defines the location and the exon-intron structure of a gene. However, it does not specify whether a gene is protein-coding and if so, where the translation start and stop are positioned. This information can be obtained by algorithms which are designed for predicting protein-coding genes in RNA transcripts, e.g., GeneMarkS-T [43] or TransDecoder [29].

The accuracy of transcriptome-based gene prediction is limited by two main factors. First, the transcriptomic evidence only covers genes which were expressed under the stud-

ied conditions. This problem can be partially mitigated by conducting multiple RNA sequencing experiments; however, this introduces a new problem: how to combine all the RNA-reads without introducing excessive noise. Second, the transcripts reconstructed from short-read RNA-Seq have been shown to be highly unreliable [44]. A comprehensive study evaluating the accuracy of transcripts obtained by mapping long RNA reads is currently in progress [45]; nevertheless, results of other studies suggest that transcripts predicted from long-read RNA sequencing are unreliable as well [46].

To offset these limitations, some gene finders combine the transcriptomic evidence with an *ab initio* component; without necessarily attempting to directly assemble/map full length transcripts. Examples of such tools, using information from splice-aligned RNA reads to improve *ab initio* predictions, are BRAKER1 [47] or GeneMark-ET [48].

2.2.3 Protein homology evidence

Protein homology-based gene prediction algorithms rely on the conservation of genes between species. Gene structures are predicted by mapping known, evolutionarily related proteins to orthologous genes in the target genome. Thus, in contrast to the genomic and transcriptomic evidence, the protein homology evidence does not originate from the target species of interest.

Protein homology gene prediction approaches are conceptually an extension to the Needleman-Wunsch [49] sequence alignment that accounts for introns by allowing long gaps with known splice site junctions. This task is commonly referred to as protein-DNA spliced alignment. Examples of protein to DNA splice aligners relying exclusively on sequence similarity are PROCRUSTES [50], exonerate [51], GenomeThreader [52], or ProSplign [53]. As was the case for transcriptomic evidence, some spliced aligners incorporate an *ab initio* component to improve their gene prediction accuracy. Examples of such algorithms are GeneWise [54], AUGUSTUS-PPX [55], or Spaln [56]. Notably, GeMoMa [57, 58] can combine protein homology with transcriptomic evidence, but lacks an *ab initio*

component.

The strength of protein homology-based approaches is the ability to transfer knowledge generated by other sequencing and annotation projects. Their weaknesses are (i) the inability to predict genes that are evolutionarily unique to the genome of interest, and (ii) a decrease in prediction accuracy with the increasing evolutionary distance between the target gene and the reference protein. This latter issue is further covered in Section 2.3.2.

The protein homology evidence is not to be confused with sequence homology, utilized by so-called comparative gene finders. These algorithms (e.g., CONTRAST [59] or AUGUSTUS-cgp [60]) exploit the conservation patterns in the alignment of multiple related genomic sequences. Because comparative gene finders are currently not frequently used (perhaps due to the difficulties associated with preparing the multiple genome alignment), the sequence homology evidence is not discussed in this thesis.

2.3 Open gene prediction challenges and opportunities for improvement

This section describes several challenges hindering the accuracy and ease of use of eukaryotic gene prediction methods and briefly outlines how we addressed them in this thesis.

2.3.1 Automatic adaptation to the diversity of eukaryotic genomes

As discussed in Section 2.2, intrinsic evidence, leveraged by *ab initio* gene prediction algorithms, plays a crucial role in gene prediction. Since eukaryotic gene organization (splicing patterns, exon length distribution, codon usage, etc.) significantly varies from organism to organism [23–26], gene finders with an *ab initio* component need to learn the species-specific properties of their target genome. Most gene prediction methods use supervised training to estimate such parameters [16], thus relying heavily on a large and high-quality set of training genes [1, 16, 17, 19, 47, 61, 62]. It has been shown that supervised algorithms trained on one species do not perform well when applied to others [13, 63]. Therefore, the high-quality training sets need to be carefully curated for each new genome of interest. The

preparation of such training sets requires manual work and validation by experts [13, 16, 17, 47]. Consequently, purely supervised training is not feasible for high-throughput, fully automated annotation. To circumvent this issue, all new algorithms described in this thesis (Chapters 3 to 5) were designed to work in a fully unsupervised manner. To enable the integration of external supervised tools (as in Chapter 4), we developed new methods that automatically prepare the necessary training sets in an unsupervised way from proteomic and/or extrinsic evidence.

2.3.2 Utilization of remote protein homologs

The accuracy of protein to DNA spliced alignment algorithms (Section 2.2.3) quickly degrades with the increasing evolutionary distance between the query DNA and the target protein [57, 63, 64]. König et al. [63] demonstrated this trend for GenomeThreader [52] and exonerate [51], two popular spliced alignment programs. They used these algorithms to predict genes in the genome of the common fruit fly (*D. melanogaster*), utilizing cross-species proteins of other flies as the gene prediction evidence. The two programs performed well (~ 0.85 exon-level F1 score) when proteins of closely related *D. simulans* were used on input. However, the accuracy rapidly decreased (F1 score ≈ 0.45) with proteins from more distant *D. grimshawi*; which is still in the same taxonomic genus. Finally, using proteins of *M. domestica*, the common house fly, further reduced the F1 score to ~ 0.4 (exonerate) and ~ 0.3 (GenomeThreader). The prediction accuracy of existing gene prediction tools combining protein homology and *ab initio* evidence (Section 2.2.3) suffers from the same issue. For example, the exon-level sensitivity of GeneWise [54], a homology-based gene predictor used in Ensembl's gene annotation system [65], was observed to be lower than 40% when the gene of interest had $< 95\%$ amino acid identity to the aligned homologous protein [54].

The steep decrease in prediction accuracy with increasing evolutionary distance of reference proteins is a serious problem because newly sequenced species often lack annotated

proteins of close-enough relatives. Chapters 3 and 4 of this thesis introduce gene prediction algorithms that were designed to utilize proteins of *any evolutionary distance*, including remote homologs. Consequently, these new algorithms can fully exploit the information contained in large and ever-growing protein databases such as OrthoDB [66, 67], EggNOG [68], or SwissProt [69]. As these databases are always readily available prior to a genome annotation project start, the tools described in Chapters 3 and 4 should be especially important for the accurate annotation of species lacking other extrinsic (e.g., transcriptomic) evidence.

2.3.3 Integration of distinct gene prediction evidence

As described in Section 2.2, gene prediction algorithms generally utilize data from one or more of the following sources: (i) intrinsic statistical patterns of the genomic sequence itself, (ii) transcriptomic evidence, and (iii) protein homology. The simultaneous utilization of all three information sources remains an open problem. The majority of tools integrating all the information (e.g., TSEBRA [70], FINDER [71], LoReAn [72], GAAP [73], IPred [74], Evigan [75], EVidenceModeler [76], JIGSAW [77], Combiner [78], or GAZE [79]) work as combiners: Their approach is to combine multiple independent *ab initio*, transcriptomic, and homology-based predictions in order to create a prediction set that is, on average, more accurate than any input source. This way, the integration of distinct information streams only occurs as a "post-processing" step of the gene prediction process. Chapter 5 describes a new approach that integrates the three data sources throughout all stages of an algorithm's training and gene prediction; thus avoiding gene prediction errors that are difficult to resolve only through the above-described combination of independent predictions.

2.4 Gene prediction accuracy metrics

Throughout this thesis, we evaluated the accuracy of gene predictions by comparing them with known gene structures contained in reference annotations of the selected target genomes. This comparison was done on two distinct levels: exon and gene. An exon was considered to be predicted correctly when both of its boundaries exactly matched the boundaries of an exon in the reference annotation. Only protein-coding exons were considered in this thesis; consequently, the outer boundaries of initial and terminal exons were defined by their translation starts and stops, respectively. The remaining exon boundaries were defined by the splice sites.

A predicted gene was considered to be correct when all its exons exactly matched *all* exons of a reference gene. If the annotation contained alternative isoforms, a gene was counted as correctly predicted when the prediction matched at least one alternative transcript. Because only a correct gene-level prediction perfectly defines the encoded protein, the evaluations in this thesis mainly focused on the gene-level accuracy.

For both exons and genes, we counted the number of correct predictions as true positives (TP), the number of incorrect predictions as false positives (FP), and the number of missed annotated exons/genes as false negatives (FN). Prediction sensitivity (Sn) and specificity (Sp) were defined as:

$$Sn = 100 \times \frac{TP}{TP + FN} \tag{2.1}$$

$$Sp = 100 \times \frac{TP}{TP + FP} \tag{2.2}$$

To combine Sn and Sp into a single measure, we computed their harmonic mean (F1):

$$F1 = 2 \times \frac{Sn \times Sp}{Sn + Sp} \tag{2.3}$$

2.4.1 Reliability of reference annotations

Comparing the predictions with known genes in a reference annotation is only meaningful when the reference annotation itself can be trusted. Unfortunately, highly reliable annotations, perfected by years of manual curation, are available only for a limited number of genomes that were subjects of pilot genome projects [3]. The annotations of three such genomes—*A. thaliana*, *C. elegans*, and *D. melanogaster*—were heavily utilized in all chapters of this thesis. To assess the prediction accuracy in genomes with less reliable annotations, the predictions were often compared with only a subset of the reference annotation. These subsets were selected to represent the most reliably annotated gene structures; the specific details of their preparation are described in each chapter.

CHAPTER 3

GENEMARK-EP+: EUKARYOTIC GENE PREDICTION WITH SELF-TRAINING IN THE SPACE OF GENES AND PROTEINS

Abstract

We present GeneMark-EP+, an unsupervised gene prediction algorithm that uses homologous cross-species proteins to guide its model training and gene prediction steps. In contrast to existing protein homology-based gene finders, which can only extract information from proteins of closely related species, GeneMark-EP+ is able to utilize proteins of any evolutionary distance, including remote homologs. Consequently, GeneMark-EP+ can fully exploit the information contained in large and ever-growing protein databases that are, unlike transcriptomic data, always readily available prior to a genome annotation project start. In tests on genomes of fungi, plants, and animals, GeneMark-EP+ delivered better prediction accuracy than *ab initio* GeneMark-ES and RNA-Seq-based GeneMark-ET, even in situations when only evolutionarily remote proteins were used on input.

3.1 Introduction

One of the major challenges of gene prediction in eukaryotes is finding an optimal way to combine sources of information extrinsic and intrinsic to the genome of interest. External information can be transferred from RNA transcripts as well as from cross-species proteins derived from annotated genomes. The integration of transcript information, e.g. RNA-Seq reads, with *ab initio* gene prediction is implemented in several algorithms and software tools, e.g. BRAKER1 [47], GeneMark-ET [48], EuGene [80, 81], and mGene.ngs [82].

The utilization of cross-species protein information is done by tools solving the problem of protein to DNA spliced alignment, e.g., GeneWise [54], GenomeThreader [52],

ProSplign [53], and Spaln [56]. Beyond a single reference protein, a reference family of homologous proteins can be used to map elements of gene structure conserved in evolution; for instance, AUGUSTUS-PPX [55] uses protein profiles derived from conserved protein domains. Information about intron position, conserved in protein primary structures of multiple homologs, was used in another tool, GeMoMa [57]. Notably, an attempt to combine protein profiles with intron position profiles for refinement of predicted genes was made by yet another method, GSA-MPSA [83].

The main weakness of methods relying on mapping homologous proteins lies in the patchiness of the evidence they generate; a sizable fraction of a whole complement of genes may code for proteins with few or no orthologs. Another weakness, as discussed in Section 2.3.2, is that protein spliced alignments become less accurate with the increasing evolutionary distance of cross-species proteins [54, 57, 63, 64]. To mitigate these issues, a protein homology-based gene finder should additionally rely on a strong *ab initio* component; this is especially important in situations when sufficient transcriptomic evidence is not available as a gene prediction input.

The application of *ab initio* algorithms for genome-wide eukaryotic gene prediction was for a long time hampered by the need for tedious and time-consuming training (Section 2.3.1; [1, 16, 17, 19, 47, 61, 62]). This issue was addressed by an *ab initio* gene finder GeneMark-ES [16, 17] which automatically estimates model parameters by iterative unsupervised training. GeneMark-ES thus does not require expert-based training or any external information for building the training set. An extension to GeneMark-ES, GeneMark-ET [48], was developed to integrate into the training process available transcript information—raw RNA-Seq reads aligned to the genome in question.

Here, we describe GeneMark-EP, an algorithm that integrates into training information extracted from a reference set of cross-species protein sequences of any evolutionary distance. To process the input protein database, we developed a specialized protein mapping pipeline called ProtHint. ProtHint first identifies a set of proteins homologous to the pro-

tein likely encoded in each putative genic locus. Then, ProtHint computes so-called *protein hints*, a set of mapped splice sites (intron borders), and translation start and stop sites along with the scores characterizing hints' confidence. The most reliable hints can be used to directly predict elements of final exon-intron structures; we call this mode of algorithm execution with direct gene structure correction GeneMark-EP+.

A key question solved by GeneMark-EP is how to find an optimal method of hint incorporation into the automatic training of an *ab initio* algorithm. Unsupervised training implemented in GeneMark-ES carries a risk of convergence to a biased set of model parameters. On the other hand, giving too much weight to protein hints may generate parameters dictated by a narrow set of conserved genes and proteins [84]. The GeneMark-EP algorithm was designed to combine strong features of both methods: (i) the ability of unsupervised iterative training of an *ab initio* gene finder to create a set of training sequences with a size beyond the reach of conventional supervised training, and (ii) the ability to correct model parameters as well as structures (the -EP+ mode) of newly discovered genes by the hints derived from homologous cross-species proteins. Thus, the new training method falls into category of gene prediction methods with semi-supervised training.

3.2 Materials

For the assessment of GeneMark-EP/EP+ as well as ProtHint accuracy, we selected annotated genomes from diverse clades: fungi, worms, plants, insects, and vertebrae (Table 3.1). The genome length varied from < 100 Mbp (*Neurospora crassa*) to > 1.3 Gbp (*Danio rerio*). With the exception of *Solanum lycopersicum*, a species representing large-genome plants important for the economy, all selected species are model organisms whose genomes presumably have a high-quality annotation. In all genomic datasets, contigs not assigned to any chromosome and the genomes of organelles were excluded from the analysis.

Table 3.1: Genomes used for assessment of GeneMark-EP and GeneMark-EP+ performance. Introns per gene values were computed with respect to the whole gene number, including single-exon genes.

Species	Assembly version	Genome size, Mb	Annotation version	# Genes in annotation	Introns per gene
Neurospora crassa	GCA_000182925	40	Broad Institute (2013)	10,785	1.7
Caenorhabditis elegans	GCA_001483305	100	WormBase WS271 (May 2019)	20,172	5.7
Arabidopsis thaliana	GCF_000001735	119	Tair Araport11 (Jun. 2016)	27,445	4.9
Drosophila melanogaster	GCA_000001215	134	FlyBase R6.18 (Jun. 2019)	13,929	4.3
Solanum lycopersicum	SL4.0	773	Consortium ITAG4.0 (Sep. 2019)	33,562	3.5
Danio rerio	GCF_000002035	1,345	Ensembl GRCz11.96 (May 2019)	25,254	8.2

Table 3.2: Characteristics of the OrthoDB v10 taxonomical space for each of the species we tested. The number of species is naturally the largest in the kingdom section of the database. *For tests in the genus-, family-, and order-excluded modes for *D. melanogaster* and *D. rerio*, the phylum was used as the largest set of reference proteins.

Number of species in the same taxonomical unit	Genus	Family	Order	Class	Phylum	Kingdom	OrthoDB root used for tests	# of proteins in the root
Neurospora crassa	0	1	7	96	364	548	Fungi	5,850,648
Caenorhabditis elegans	2	2	4	5	6	447	Metazoa	8,266,016
Arabidopsis thaliana	1	7	9	-	99	116	Plantae	3,510,742
*Drosophila melanogaster**	19	19	55	147	169	447	Metazoa	8,266,016
Solanum lycopersicum	1	9	10	-	99	116	Plantae	3,510,742
*Danio rerio**	0	4	4	49	245	447	Metazoa	8,266,016

3.2.1 Protein database preparation

We used OrthoDB v10 protein database [66] as an all-inclusive source of protein sequences. Still, for generating protein hints for a particular species, we used subsets of OrthoDB: plant proteins for gene prediction in *Arabidopsis thaliana*, arthropod proteins for gene prediction in *Drosophila melanogaster*, etc. (Table 3.2).

A principal feature of ProtHint and GeneMark-EP+ is their ability to extract information from multiple homologous proteins. To evaluate this ability, we had to model practical situations when the evolutionarily distance between the genome of interest and the most closely related species with a sequenced and annotated genome varies significantly. To simulate these variations in our tests, we introduced restrictions on evolutionarily distance to the closest species from which the target proteins could be recruited. These restrictions were implemented by removing from the protein database (i) proteins encoded in the genome of a given species; (ii) proteins from all species from the same subgenus; (iii) pro-

teins from the same genus; (iv) proteins from the same family; (v) proteins from the same order; and (vi) proteins from the same phylum. Notably, distributions of numbers of species within a genus, family, etc. defined by a given species are species-specific (Table 3.2).

3.2.2 APPRIS principal isoform annotation

As an additional test set, we used annotation of major protein isoforms available in the APPRIS database [85]; this assessment was done for *C. elegans*, *D. melanogaster*, and *D. rerio*. Arguably, the accuracy of prediction of major isoforms is of significant interest, since in most gene loci, the major isoform was observed to be expressed in significantly higher volume than other (minor) isoforms [86].

3.3 Methods

3.3.1 Overview of GeneMark-EP/EP+

GeneMark-EP/EP+ executes the following tasks: (i) selecting genomic regions, *seed regions*, containing gene candidates (*seed genes*); (ii) identifying a set of homologous proteins for each seed gene; (iii) constructing spliced alignments of homologous proteins to each seed region and generating hints to exon-intron structures; (iv) running iterative semi-supervised training; and (v) making the final gene prediction without (EP mode) or with an additional option (EP+ mode) to enforce high-confidence protein hints in the predicted exon-intron structures.

Tasks (i)-(iii) are devoted to generating protein hints and are solved by the ProtHint pipeline. Tasks (iv) and (v) correspond to the training and prediction steps of GeneMark-EP and -EP+. At these steps, we use the hints to exon-intron structure coordinates as an input to an expectation-maximization-type algorithm that finds models of compositional patterns of protein-coding and non-coding regions simultaneously with the most likely parse of genomic sequence into coding and non-coding regions.

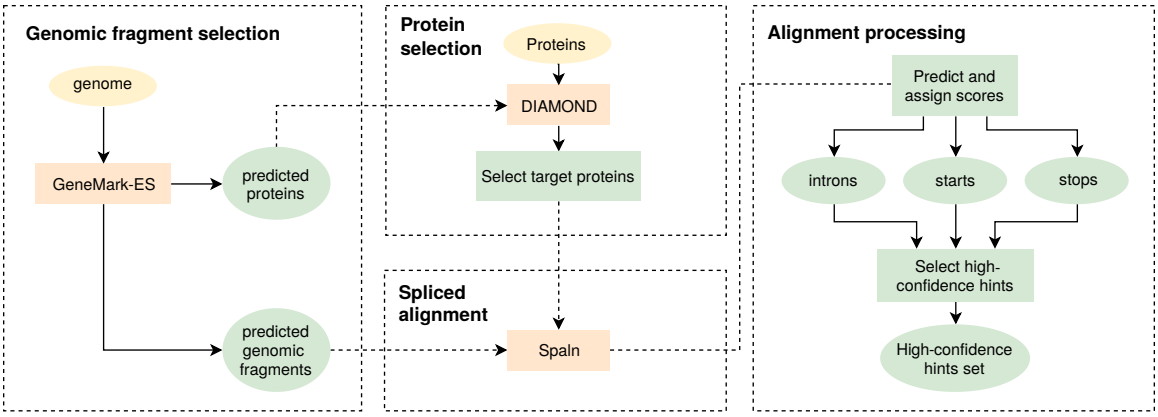

Figure 3.1: An overview of the ProtHint pipeline.

3.3.2 ProtHint: generating protein hints from a large protein database

3.3.2.1 ProtHint's general logic

The role of the ProtHint protein mapping pipeline is to accurately predict locations of exon boundaries (so-called protein hints) from a large number of proteins of any evolutionary distance to the genome of interest. The workflow of ProtHint (Figure 3.1) is as follows.

ProtHint starts by running the unsupervised gene finder GeneMark-ES [16] to define candidate *seed regions* containing putative protein-coding genes (*seed genes*). Each seed gene is translated to a protein and queried against the entire input protein database (e.g., OrthoDB) using DIAMOND [87] in the BLASTp mode. A set of proteins with statistically significant hits ($e\text{-}val < 1e{-}3$) defines target proteins presumed to be homologous to the seed gene query. The best-scoring target proteins (up to 25 per seed) are splice aligned back to the seed region (extended by 2000 nt both upstream and downstream) with Spaln [56].

The resulting alignments are processed to predict the locations of exon boundaries (i.e., hints to coordinates of introns and translation start/stop codons). ProtHint scores each such hint to remove spurious alignments and to classify the reported predictions into low- and high-confidence groups. The details of the scoring system are described in the following sections. By focusing on individual exon boundaries, instead of trying to infer the full gene

structure from each protein alignment, ProtHint can predict accurate hints from conserved domains of otherwise remotely related proteins.

3.3.2.2 Score system for introns

The evolutionary conservation between primary structures of target proteins and a protein encoded in the seed region has to be quantified to evaluate the reliability of introns predicted from spliced alignment. To facilitate this quantification, we define three types of scores.

Alignment of an Entire Exon (AEE) score is defined as a score of the Spaln (or ProSplign) alignment of an exon translation and a target protein. The AEE scores are computed for all exons adjacent to introns mapped by spliced alignments. The alignment score is computed with BLOSUM62 [88] substitution parameters and a linear gap penalty $= -4$. The AEE score is not normalized by the exon length; therefore, exons with low scores are either too short or they are long and poorly aligned. At the initial step of the algorithm, we keep introns bordered by exons with high AEE scores (further described in Section 3.3.2.3).

Intron Borders Alignment (IBA) score characterizes the conservation of exons adjacent to the scored intron, with larger weights given to parts close to the donor and acceptor splice sites. First, scores S_d and S_u are computed for the downstream and upstream exons (relative to the intron) defined by the spliced alignment. S_d is defined as:

$$S_d = \sum_{i=1}^{w} S_a(G_i, P_i) \times W(i) \tag{3.1}$$

Here, $S_a(G_i, P_i)$ is a BLOSUM62 [88] substitution score defined for a target protein amino acid P_i and a codon-defined amino acid G_i. Gaps are penalized with a linear gap penalty

$= -4$. w defines a window width (by default, $w = 10$) and $W(i)$ is a weight function,

$$W(i) = \frac{K(i)}{\sum_{i=1}^{w} K(i)} \tag{3.2}$$

where $K(i)$ is a kernel value for position i counting in codons from the acceptor site. For instance, a linear kernel (the default kernel) defines this value as:

$$K(i) = 1 - \frac{|i| - 1}{w} \tag{3.3}$$

A value IBA_{raw} is a geometric mean of S_d and S_u (S_u is computed in the same way as S_d):

$$IBA_{raw} = \begin{cases} \sqrt{S_d \times S_u}, & \text{if } \min(S_d, S_u) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Finally, the IBA score is obtained by normalizing the IBA_{raw} score into a $[0, 1]$ range:

$$IBA = \frac{IBA_{raw}}{\max(S_a)} \tag{3.5}$$

where $\max(S_a)$ is a maximum score among the elements of the amino acid substitution matrix.

Intron Mapping Coverage (IMC) score is a count of how many times a given intron was exactly mapped by spliced alignments of distinct target proteins. The IMC score is computed only from the set of introns with IBA score > 0.1; in order to prevent high scores through the accumulation of noise (details in Section 3.3.2.3).

3.3.2.3 Application of the intron scores

ProtHint uses the following method to filter scored introns. First, introns whose two adjacent exons have AEE scores $\geq E_t$ are selected; where E_t is a chosen threshold. For

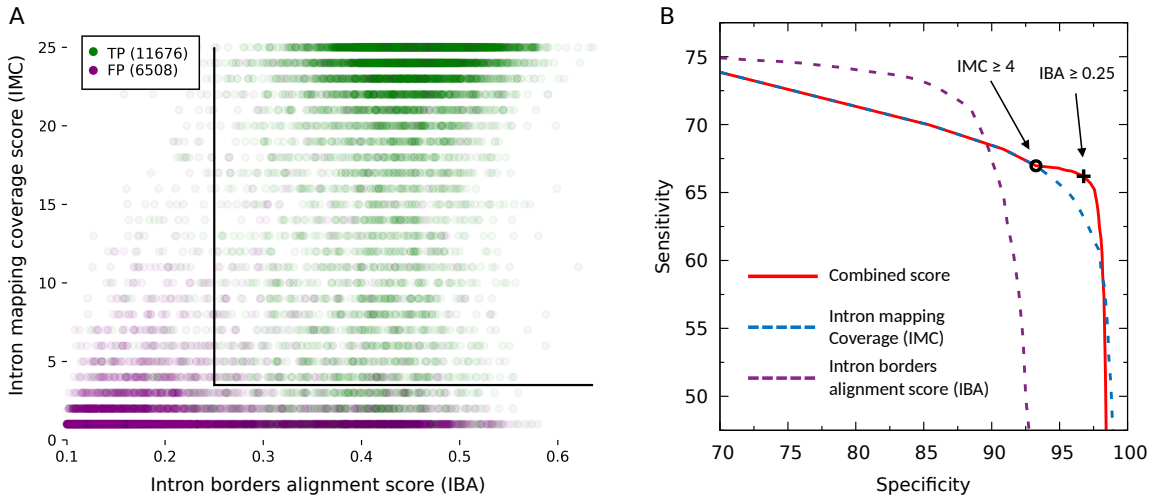

Figure 3.2: ProtHint high-confidence intron processing, shown for *N. crassa*. Introns were generated by spliced alignments of target proteins from species beyond *Neurospora* genus. (A) Distribution of the score vectors (IBA, IMC) of true positive (green) and false positive (purple) introns. The black lines represent cutoffs at IMC $= 4$ and IBA $= 0.25$. Total numbers of false and true positives are shown in the upper left corner. (B) Sn and Sp of intron sets selected by thresholds on IBA score and IMC score. IMC score is computed for introns that have IBA score ≥ 0.1 and exon AEE score ≥ 25. The red curve represents the following. The left branch of the curve reflects (Sp, Sn) values of the sets of introns selected by shifting the IMC threshold from 0 to 4. The one with the IMC threshold = 4 is recorded as set A – the set corresponding to the black circle in the red curve. Then, the right branch of the curve reflects (Sp, Sn) of the set of introns generated by applying to set A the IBA score threshold changing from 0 to 0.25 and up to 1.0. Set B corresponds to the black cross in the red curve, introns in this set have IMC ≥ 4 and IBA ≥ 0.25. Separate curves for IMC score change (dashed blue) and IBA score change (dashed purple) are shown as well.

$E_t = 25$, in modeling on known genomes, we observed a relatively high Sn value of the candidate introns (Figure A.1). Further increase of E_t eliminated true introns while not significantly improving the Sp value (Figure A.1). Thus, E_t defaults to 25.

Next, a subset with an IBA score $\geq I_t$ is selected; where I_t is another chosen threshold. Our modeling has shown an increase in the Sp value of the candidate introns for $I_t = 0.1$ that occurred without a noticeable change in the Sn (Figure A.1). Thus identified subset of introns represents a set of all mapped introns; this is used as an external evidence to generate anchored introns for the GeneMark-EP training steps (described in Section 3.3.3.1).

Finally, within the set of all mapped introns, ProtHint selects a narrower set of *high-*

confidence introns. At this stage, introns defined by alignments of multiple distinct target proteins are collapsed to compute the IMC score. The IBA score of the collapsed introns is defined as a maximum of the individual IBA scores. High-confidence introns must have canonical GT-AG splice sites, an IMC score ≥ 4 and an IBA score ≥ 0.25 (Figure 3.2). GeneMark-EP uses the high-confidence introns to estimate the initial parameters of its intron model. Further, these introns are enforced in the prediction step of the GeneMark-EP+ mode (Section 3.3.3.2).

3.3.2.4 Score system for translation starts and stops

Similarly to scores introduced for intron hint generation, we define a Border Alignment Quality (BAQ) score for translation starts and stops. This score is computed for w amino acids downstream (upstream) of a start (stop) codon, weighted by a kernel-dependent function (Equation (3.1)).

Next, a Site Mapping Coverage (SMC) score counts the number of N-terminals (C-terminals) of distinct target proteins aligned to a particular start (stop) codon position of a candidate gene. The SMC scores are computed only from the sets of initial (terminal) exons whose BAQ scores are > 0.

Start codons are additionally scored by counting the number of protein alignments overlapping a given start. A precise definition of the overlap is as follows. Start S is considered to be overlapped by a target protein P if an exon E in P overlaps S upon spliced alignment. Still, to be counted as overlapping, exon E needs to satisfy these criteria: (i) AEE score of E has to be ≥ 25. (ii) The spliced alignment of protein P must contain a mapped start codon or an acceptor site (within the set of all reported starts/introns) that coincides with the exon start. In other words, the start of the overlapping exon must define either a start codon or an acceptor splice site (Figure 3.3).

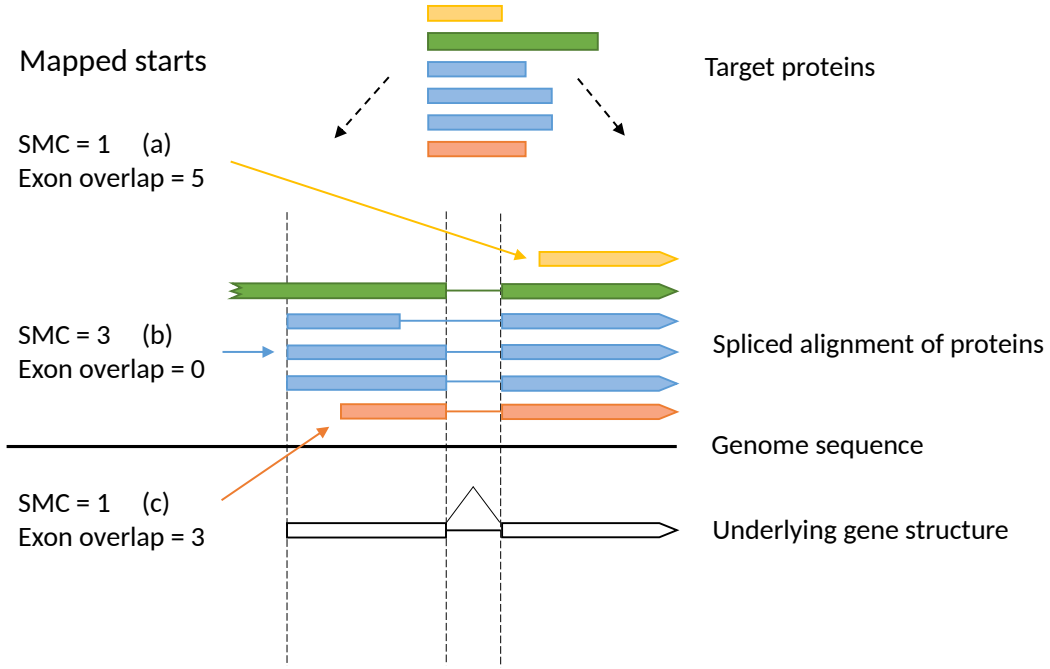

Figure 3.3: Gene start mapping coverage (SMC) scores and counts of exon overlaps. Start (a) is overlapped by five exons that coincide with an upstream intron. Start (b) is overlapped by one exon (green) but this exon's upstream boundary does not coincide with an end of an intron or a start codon mapped by ProtHint, therefore it does not contribute to the exon overlap. Start (c) is overlapped by three exons which define an upstream start, green exon is again not counted.

Table 3.3: Sensitivity and specificity of all gene start hints created by ProtHint as well as of the high-confidence start hints. High specificity was achieved with filtering by SMC scores as well as by the removal of candidate starts overlapped by at least one target protein (suggesting that a start is located upstream). Sn was defined with respect to a full complement of starts, including alternative ones as given in annotation. The numbers were generated in tests with reference proteins from species outside the relevant genus. Results for all test species are shown in Table A.1.

		All reported starts	Filtered with SMC >= 4	Filtered with SMC >= 4 and exon overlap =0
A. thaliana	Sn	69.3	62.9	61.4
	Sp	70.9	89.8	94.4

3.3.2.5 Application of the start and stop scores

Translation starts and stops are filtered as follows. A start codon candidate is an ATG codon present in a mapped initial exon and aligned to an N-terminal methionine in the target protein; a stop codon candidate is a stop codon in a mapped terminal exon. The initial (terminal) exon containing a candidate gene start (stop) must have AEE score ≥ 25; otherwise the candidate start (stop) is removed. ProtHint subsequently removes starts (stops) with BAQ score $= 0$.

To select a subset of high-confidence hints, ProtHint chooses stop codon candidates with SMC score ≥ 4 as well as start codon candidates with SMC score ≥ 4 and no overlap by longer target proteins (Figure 3.3). The set of high-confidence hints to translation starts and stops is used to estimate parameters of GeneMark-EP models of translation initiation and termination sites. Also, the high-confidence hints are directly enforced in the prediction step of GeneMark-EP+.

Tables A.1 and 3.3 illustrate how the application of these rules leads to an increase in the specificity of the predicted starts.

3.3.3 Integration of genomic sequence patterns and protein homology into gene prediction

3.3.3.1 Model training

The iterative training of GeneMark-EP's statistical models (Figure 3.4) works as follows. In the first iteration, full-length introns mapped by ProtHint with scores exceeding a stringent threshold (high-confidence hints) are used to estimate parameters of splice site models as well as branch point site models (particularly important for intron models of fungal genomes). The splice site models together with heuristic models of protein-coding and non-coding regions make a complete set of models of a generalized hidden Markov model (GHMM) [16]. The models are used in the first run of the Viterbi algorithm (see [89]) that generates a maximum likely parse of genomic sequence into coding and non-coding regions—the parse delineating the first set of genes predicted by GeneMark-EP. Next, GeneMark-EP analyzes available data to make updated training sets and re-estimate model parameters. Coordinates of exons predicted by GHMM are compared with intron hints determined by ProtHint. This comparison leads to the selection of so-called *anchored* elements—exons with at least one splice site identified by both GHMM and ProtHint. A set of the anchored exons along with a set of predicted single-exon genes (with length > 800 nt) comprises an updated training set for a three-periodic Markov chain model of protein-coding regions [90]. Sequences of introns bounded by two anchored splice sites as well as intergenic sequences bordered by anchored terminal and initial exons of adjacent genes (Figure 3.5) are used to update parameters of the non-coding region model. The set of all updated models is used by the Viterbi algorithm to generate a new set of predicted genes. A new update of anchored elements and the next round of parameter re-estimation follows.

Several probability distributions used in GeneMark-EP, such as length distributions of exon, intron, and intergenic regions, are initially defined as uniform ones. A more accurate estimation of these distributions is done in subsequent steps of the iterative training (Figure 3.4). Prior to the final iteration, the estimates of the GHMM transition probabilities,

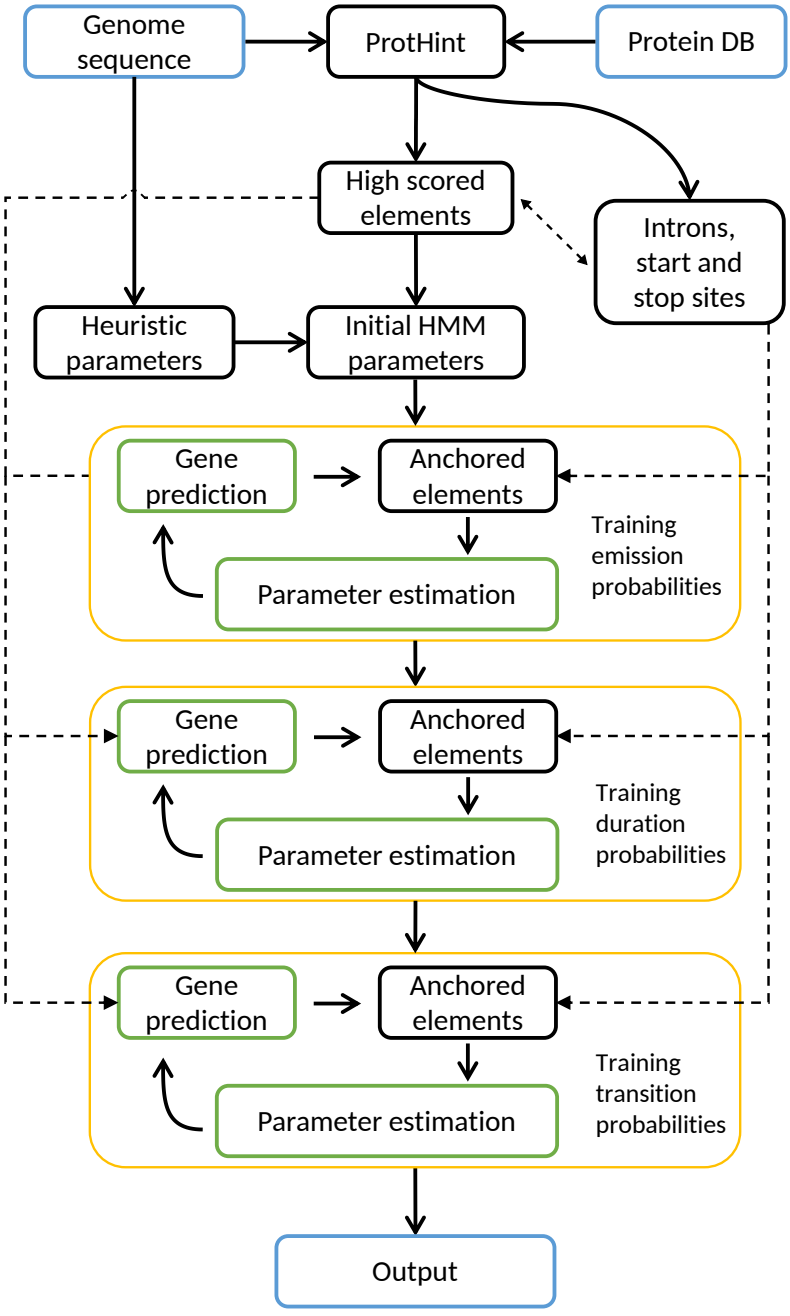

Figure 3.4: A flowchart of the GeneMark-EP, EP+ iterative training.

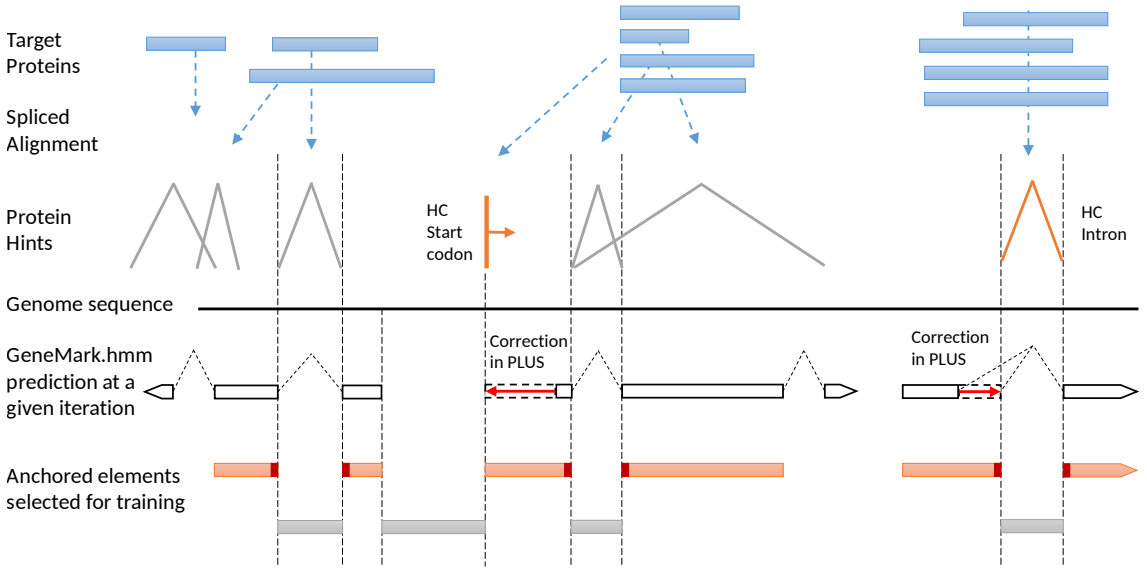

Figure 3.5: Selection of anchored elements for GeneMark-EP+ training with enforcement of High-Confidence (HC) hints.

affecting the frequencies of genes with a given number of introns, are estimated. Also, the parameters of three-phase models of splice sites indexed by a nucleotide position after which the intron divides a codon triplet are only estimated in the final steps. In experimental runs for genomes of different lengths, we have verified that six iterations are sufficient for GeneMark-EP to reach convergence in coordinates of predicted genes and the values of model parameters.

3.3.3.2 Final gene prediction

Gene predictions made in the final iteration are reported as the output of GeneMark-EP. The Viterbi algorithm can also be run with enforcement of high-confidence elements mapped by ProtHint. Particularly, this is done by modifying components of an objective function of the Viterbi algorithm associated with chosen hidden states. The sites that are enforced receive high values of the objective function to ensure their addition to a path selected by the optimization algorithm seeking the maximum value of the log Viterbi objective function. This mode of gene prediction produces the GeneMark-EP+ output.

Note that GeneMark-EP/EP+ algorithms are designed to predict non-overlapping genes

with no alternative isoforms. This design suits the paradigm that each gene locus encodes a major (expressed in most tissues) protein isoform [85].

3.3.3.3 General updates to the GHMM architecture

Compared to the published versions of GeneMark-ES and -ET, several updates to the underlying GeneMark's GHMM architecture were made. Since these updates improve the general gene prediction accuracy of all algorithms in the GeneMark family; they were also added to GeneMark-ES and -ET.

The first change involves updates to the modeling of intergenic length distribution. Instead of using a uniform distribution with a fixed maximum length, intergenic regions are now modeled using a non-parametric estimation of a probability density function. The non-parametric estimation is applied in the last (third) round of self-training iterations (see flowchart in Figure 3.4). Furthermore, uniformly distributed pseudocounts are added to smooth the distributions. In GeneMark-EP and ET, only those intergenic regions that are situated between genes with anchored introns are used as data for the non-parametric estimation of the length distribution.

The default minimum length of genes predicted by GeneMark-ES, -ET was set to 300 nt. Currently, shorter genes are allowed in the final prediction when supported by extrinsic hints.

Finally, GeneMark-EP introduced a model for non-canonical introns with GC-AG splice sites. The prior probability of GC-AG introns (compared to the canonical GT-AG ones) is set to 0.001.

3.3.4 Methods related to algorithm assessment

This section describes the design of methods that were used to evaluate various questions about the performance of GeneMark-EP/EP+ and ProtHint; other than the standard accuracy assessment described in Section 2.4. In all evaluations, regions of annotated pseudo-

genes were excluded from comparisons.

3.3.4.1 Repeat masking

All tests were run on repeat-masked genomes. Repetitive sequences (interspersed repeats and low complexity sequences) were identified by RepeatModeler [91] and RepeatMasker [92]. A run of RepeatModeler on a whole genome produced a repeat library. Next, the locations of repeats were identified and soft-masked by RepeatMasker.

3.3.4.2 Assessment of genomes with unreliable reference annotation

To assess the accuracy of gene prediction made for *A. thaliana*, *C. elegans*, *D. melanogaster*, and *N. crassa*, we compared genes predicted and annotated on a whole genome scale. In the case of *S. lycopersicum*, we made an additional sensitivity evaluation using a limited set of genes that had all introns supported by RNA-Seq mapping. The RNA-Seq data was mapped and sampled from NCBI's Sequence Read Archive [93] by VARUS [94]. In the case of *D. rerio*, we excluded annotated partial exons (ubiquitous in this genome) from exon-level accuracy assessment; we computed gene-level sensitivity only for genes having in annotation complete alternative transcripts.

3.3.4.3 The effect of using partially mapped proteins

ProtHint scores and filters all protein hints individually, irrespective of the quality of the global protein alignment. As a result, it can extract accurate hints from conserved domains of otherwise evolutionarily distant proteins. To evaluate the extent to which these "partial" hints affect the final gene prediction accuracy, we ran GeneMark-EP+ with high-confidence introns exclusively supporting full gene structures (determined by comparing the mapped hints with annotation) and compared this run with a GeneMark-EP+ run using all mapped high-confidence introns. Notably, because we selected the set of "complete" hints based on a comparison with the reference annotation, this set did not contain any false positives.

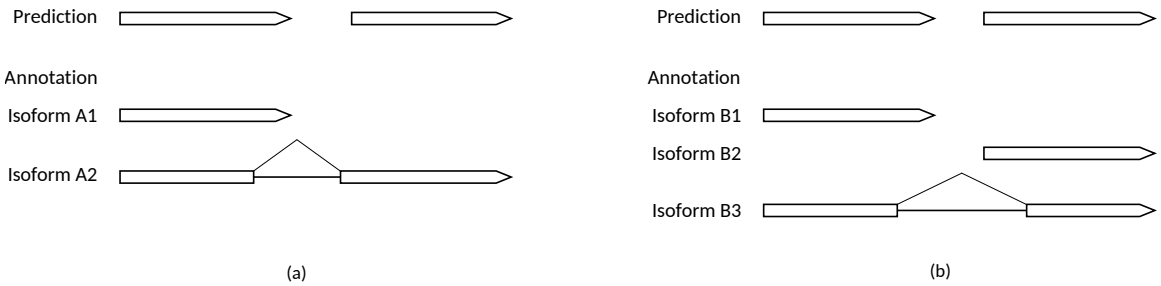

Figure 3.6: Gene splitting events caused by alternative isoforms that include other isoforms as their components. We removed such cases from the test set for gene splitting assessment. (a) Isoform A1 is correctly predicted. As a result, full isoform A2 cannot be predicted at the same time and it is split. (b) The algorithm makes correct predictions of isoforms B1 and B2. If isoform B3 was considered as the annotation, it would be split by the prediction.

3.3.4.4 Assessment of gene merging and gene splitting errors

Gene merging and splitting errors are expected to be reduced by the use of protein hints to gene translation starts and stops. This expected improvement in prediction accuracy of GeneMark-EP+ can be more accurately observed on properly prepared test sets. Prior to the evaluation of gene splitting, we excluded from the test sets (i) genes fully overlapping shorter genes present inside introns in any strand; (ii) genes with larger isoforms combining or including shorter alternative components (Figure 3.6); and (iii) genes with introns longer than 10,000 nt (the default maximum intron length). For genes with annotated multiple alternative isoforms, we used the longest one as a representative. Prior to the evaluation of gene merging, overlapping genes present in annotation (e.g., a gene within an intron) were merged into a single gene in order to exclude such cases from being counted as merged genes.

3.3.4.5 Do introns mapped by ProtHint tend to occur in gene regions coding for conserved domains?

To address this question, we used the following procedure. Annotated genes were translated to proteins and used as queries in RPS-BLAST [95] to search (*e-val* $< 1e{-}2$) against NCBI's Conserved Domains Database [96]. The results of the RPS-BLAST searches were

processed with a rpsbproc utility [96] to generate a map of conserved domains for each RPS-BLAST query. Finally, coordinates of the conserved domains were mapped back to the seed region of genomic DNA and compared with the ProtHint output to find out how many introns were mapped into regions coding for conserved domains. We conducted this analysis for genes of *D. melanogaster*, *C. elegans*, and *D. rerio*—genomes annotated in the APPRIS database [85] representing genes coding for principal protein isoforms.

3.4 Results

We compared the gene prediction accuracy of GeneMark-EP and -EP+ with the accuracy of GeneMark-ES and GeneMark-ET. In addition, we made a detailed accuracy assessment of ProtHint. The assessments were done using the genomes of six species: *N. crassa*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *S. lycopersicum*, and *D. rerio* (see Section 3.2); using the accuracy evaluation methods presented in Section 3.3.4. Importantly, we executed ProtHint and GeneMark-EP/EP+ with multiple protein databases on input, to simulate different evolutionary distances of available input proteins (see Section 3.2.1).

3.4.1 Accuracy assessment of GeneMark-EP+ and comparison with GeneMark-ES

We first present the final results of GeneMark-EP+ (with the enforcement of high-confidence hints) since it proved to be more accurate than GeneMark-EP (without the enforcement) in all tested scenarios. Details on the accuracy assessment of GeneMark-EP vs GeneMark-EP+ are given in Section 3.4.3.

For each genome (Table 3.1), we determined how the accuracy of GeneMark-EP+ at the *gene level* (Figure 3.7) and *exon level* (Figure A.2) depended on the choice of a set of the reference proteins. The pattern of accuracy changes at the gene level was similar to the one observed at the exon level; therefore, we show the results of the accuracy assessment at the gene level in the main text, while the results for the exon level are provided in the Appendix.

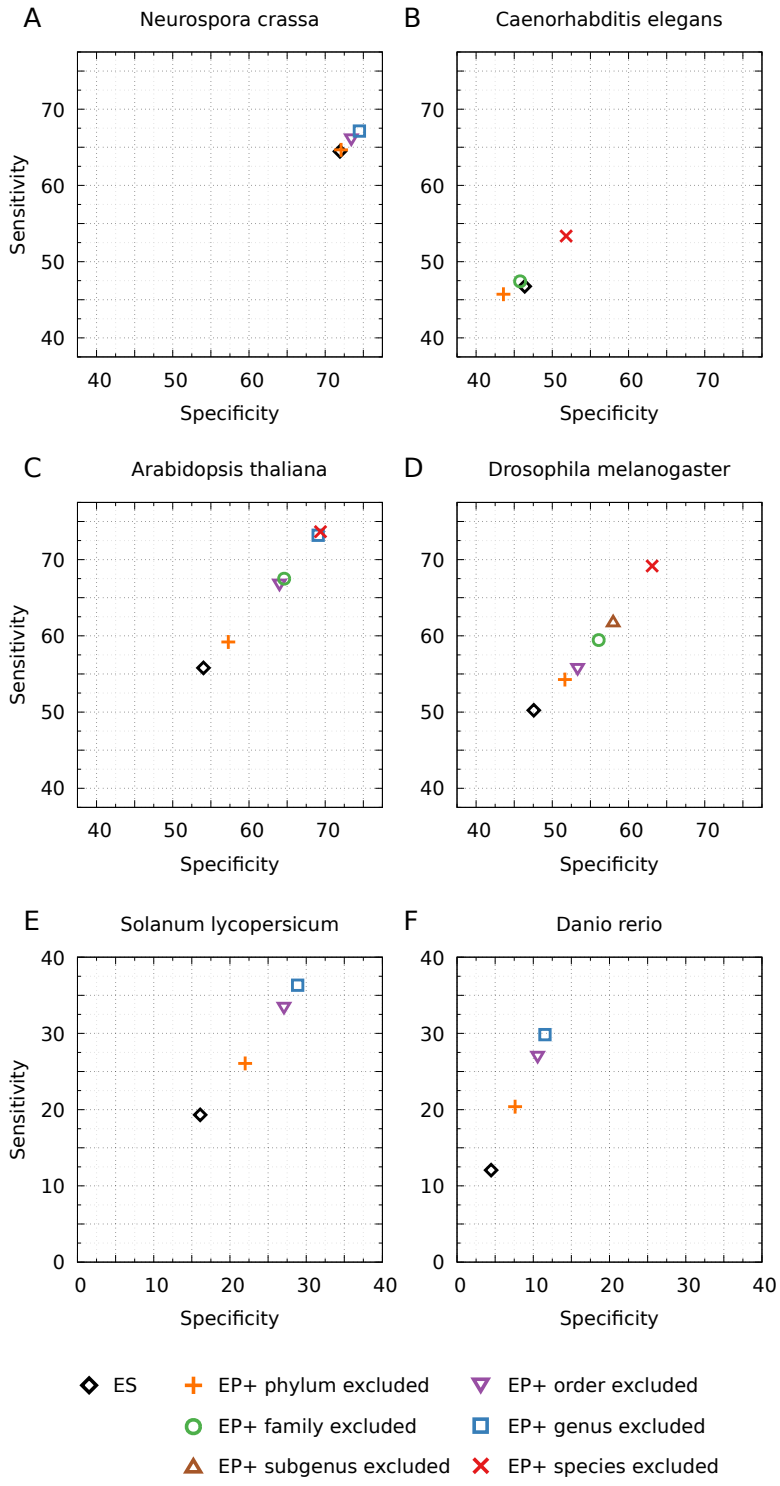

Figure 3.7: A comparison of GeneMark-ES and GeneMark-EP+ accuracy on the gene level. The accuracy of GeneMark-EP+ is shown for cases when ProtHint works with different size sets of reference OrthoDB proteins: from the largest (only proteins from the same species are excluded) to the smallest (proteins of the whole phylum excluded).

We present the results separately for three groups of genomes: fungal genomes, compact eukaryotic genomes and large eukaryotic genomes.

3.4.1.1 Fungal genomes (N. crassa)

The accuracy of GeneMark-ES was high, as it has been typical for fungal genomes ([17]). Even with hints originating from the largest set of reference proteins, those outside of genus or order, GeneMark-EP+ improved the Sn value of GeneMark-ES only by $\sim$2 percentage points (Figure 3.7A). With a smaller set of more remote reference proteins, originating from the species outside of the fungal phylum, the accuracy of GeneMark-EP+ matched the accuracy of GeneMark-ES (Figure 3.7A). This result went in line with previous observations showing that GeneMark-ES is a highly efficient *ab initio* gene finder for fungal genomes [17].

3.4.1.2 Compact eukaryotic genomes (C. elegans, A. thaliana, and D. melanogaster)

When GeneMark-EP+ used the largest set of reference proteins (just without proteins from the same species), for *A. thaliana* and *D. melanogaster*, we observed an improvement of $\sim$20 percentage points in both Sn and Sp in comparison with GeneMark-ES (Figure 3.7CD). As the target proteins were coming from larger and larger evolutionary distances, the accuracy did steadily decrease. When the target proteins were selected from outside of the same phylum, there was an increase of only 5 percentage points in gene-level Sn and Sp in comparison with GeneMark-ES.

For *C. elegans*, when the set of reference proteins excluded just proteins of the same species, GeneMark-EP+ improved the accuracy of GeneMark-ES by $\sim$6 percentage points (Figure 3.7B). We observed almost no difference between GeneMark-EP+ and GeneMark-ES when the reference proteins were only from species outside the *C. elegans* family and a slight decrease in accuracy ($\sim$2 percentage points) for reference proteins outside of the taxonomical phylum. Notably, the gene-level accuracy for *C. elegans* was lower than that

Table 3.4: An accuracy assessment for *S. lycopersicum*. Only genes which had all introns in the gene supported by RNA-Seq mapping were selected for the test set A. All the other genes were selected into set B. Single-exon genes were excluded from this analysis. Set A contained 15,832 genes with 84,424 introns. Set B contained 9,506 genes with 34,282 introns.

GeneMark-	ES		EP+ genus excl.		EP+ order excl.		EP+ phylum excl.	
Test Set	**A**	B	**A**	B	**A**	B	**A**	B
Gene Sn	**22.7**	4.8	**53.0**	8.4	**47.8**	8.0	**34.0**	6.9
Exon Sn	**76.5**	46.1	**88.8**	52.5	**87.3**	51.9	**81.5**	48.7
Intron Sn	**79.5**	53.3	**93.9**	61.7	**92.7**	60.9	**86.3**	56.8

ProtHint HC	genus excluded		order excluded		phylum excluded	
Test Set	**A**	B	**A**	B	**A**	B
Intron Sn	**87.6**	50.1	**79.8**	43.8	**30.3**	14.3
Start Sn	**60.4**	21.5	**48.2**	15.2	**5.6**	1.4
Stop Sn	**69.3**	20.5	**54.3**	14.7	**8.1**	1.7

for other species with compact genomes.

3.4.1.3 Large eukaryotic genomes (S. lycopersicum and D. rerio)

The gene-level accuracy of GeneMark-ES was low for these genomes (between 5 and 20 percentage points). GeneMark-EP+ improved the accuracy for *S. lycopersicum* by $\sim$15 percentage points when it used a protein reference set from species outside of the tomato genus or order (Figure 3.7E). For *D. rerio*, having a reference set of proteins without those from the same genus or the same order led to Sn and Sp improvements of $\sim$20 and $\sim$5 percentage points, respectively (Figure 3.7F). However, the improvements were twice as low when reference proteins were available only outside of the *S. lycopersicum* or *D. rerio* phyla.

The relatively low gene prediction accuracy in large genomes could be partially attributed to incorrect and/or incomplete gene annotations. Therefore, we made an additional effort to refine the test sets in *S. lycopersicum* and *D. rerio* by selecting genes supported by RNA-Seq data and complete genes, respectively (Section 3.3.4.2).

We observed that annotated genes of the *S. lycopersicum* genome supported by RNA-Seq hints were significantly better predicted by GeneMark-EP+ (Table 3.4). We divided the

annotated multi-exon genes into two groups: (a) genes with all introns supported by RNA-Seq and (b) all other genes. GeneMark-EP+'s sensitivity (for a GeneMark-EP+ run having reference proteins outside of the *S. lycopersicum* genus) was better by 40 percentage points in the set (a) compared to the set (b) on the gene, exon, and intron levels. It is important to emphasize that RNA-Seq information was not used in GeneMark-EP+. Sensitivity defined for the set of introns mapped by ProtHint was also better in the set (a) by $\sim$40 percentage points (Table 3.4).

As already mentioned, the annotation of *D. rerio* contained many partial exons that in turn were parts of incomplete transcripts. We evaluated the exon-level Sn separately for exons within complete and incomplete transcripts (Table A.3) and observed a 75.1% exon Sn in the "complete" group versus 67.6% in the "incomplete" group. Similarly, gene-level sensitivity was better by 6 percentage points in predicting genes with complete transcripts compared to all genes (Table A.3).

3.4.1.4 Summary for all groups

Altogether, we observed that for the majority of the considered species, the accuracy of GeneMark-EP+ was better than the accuracy of GeneMark-ES, regardless of how large a set of reference proteins was used for spliced alignments. For the fungal genome, *N. crassa*, an improvement was negligible due to the ability of GeneMark-ES to deliver high accuracy for fungal genomes; we also observed a small decrease of accuracy in the *C. elegans* test with a phylum-excluded reference set of proteins (addressed in the Discussion).

3.4.2 Accuracy assessment of ProtHint

The main role of ProtHint is the generation of coordinates (and their confidence scores) of potential borders between coding and non-coding regions in a novel genome. Specific thresholds on confidence scores are defined to select different subsets of hints (e.g., the high-confidence set). The GeneMark-EP training procedure can tolerate a high number

Table 3.5: Accuracy of ProtHint for the *D. melanogaster* genome: sensitivity and specificity of hints to introns, start and stop codons. The results are shown for all reported hints or just high-confidence hints. Results for all tested species are shown in Table A.4.

	The level of exclusion of database proteins									
	Species		Subgenus		Family		Order		Phylum	
D. mel.	All reported	High conf.	All reported	High conf.	All reported	High conf.	All reported	High conf.	All reported	High conf.
Intron Sn	79.8	74.6	72.8	62.6	66.2	54.3	49.7	34.4	35.8	20.9
Intron Sp	83.5	**98.9**	79.6	**98.8**	79.5	**98.8**	80.5	**99.0**	88.4	**99.5**
Start Sn	70.3	60.7	49.8	36.5	37.7	29.2	22.3	15.9	14.1	9.7
Start Sp	79.5	**97.4**	75.6	**96.7**	71.6	**95.6**	73.4	**94.5**	75.0	**93.5**
Stop Sn	75.3	68.4	56.7	45.2	44.7	36.9	26.7	19.8	15.8	11.2
Stop Sp	94.8	**99.3**	94.2	**98.8**	92.8	**98.5**	94.5	**98.9**	95.8	**99.2**

of false positive intron hints since only a subset, the anchored introns, is used in training (Section 3.3.3.1). Thus, the set of all mapped hints should have high Sn while the Sp level can be lower. On the other hand, the high-confidence hints—those utilized in the initial GeneMark-EP parameter estimation as well as in the hints' enforcement—must have high Sp, as these hints are directly included in the final gene predictions.

3.4.2.1 Sensitivity of all protein hints

When the maximum set of reference proteins was used (all proteins except those from the same species), the set of intron hints generated by ProtHint had $> 75\%$ intron Sn and $\sim 70\%$ Sn for translation starts and stops (Tables A.4 and 3.5). The Sn was dropping down steadily as the evolutionary distance to reference proteins was increasing. Particularly, when the proteins from species of the same order were excluded, the Sn was, on average, $\sim 65\%$ for intron hints and $\sim 40\%$ for translation start and stop hints. The largest reduction in the volume of the protein reference set—the exclusion of proteins from the same phylum—decreased Sn of all reported intron hints down to $\sim 40\%$ on average (Table A.4). Here, the largest Sn value (the fraction of correct intron hints) was observed for *N. crassa* (60%), and the lowest one for *C. elegans* (26%). For the same protein set, the Sn of translation start and stop hints varied significantly between the species, from 8% for *C. elegans* to 30% for *N. crassa* (Table A.4).

3.4.2.2 Specificity of high-confidence protein hints

The sets of high-confidence hints were observed to have high specificity, averaging over 95% (i.e., 5% of false positives) over the six species (Table A.4). This level remained high even for the smallest sets of reference proteins, proteins from species outside of the phylum of interest (Tables A.4 and 3.5). In the case of *C. elegans*, along with high Sp, we observed a low Sn value of the high-confidence hints, which can be explained by the presence of just a few species with sequenced genomes in the *C. elegans* phylum (Table 3.2). For all other species, when compared to the simultaneous increase in Sp, a decrease in Sn upon transition from all mapped to high-confidence hints was small (Tables A.4 and 3.5).

3.4.2.3 ProtHint results with non-default settings

All presented ProtHint results were shown for ProtHint runs with the default parameter settings. This section describes results showing how changes in several important ProtHint parameters affect the prediction accuracy. The significance of these results and their impact on the ProtHint design is fully discussed in the Discussion section.

Choice of thresholds for high-confidence filtering Figure 3.2A shows the distributions of vectors representing intron hints generated for *N. crassa* (both false and true as compared with annotation). Figure 3.2B shows the corresponding Sp-Sn curves; generated for sets of intron hints obtained by filtering with changing IMC and IBA thresholds. The distribution of the score vectors (Figure 3.2A) as well as the behavior of the Sp-Sn curves (Figure 3.2B) depends on the selection of the set of reference proteins (genus or order or phylum excluded; Figures A.3 to A.5). A choice of thresholds selecting high-confidence intron hints eventually affects the accuracy of GeneMark-EP+. We assessed the extent of this effect for *A. thaliana*, *N. crassa*, and *S. lycopersicum* (Figures A.3 to A.5). It was shown that the best average prediction accuracy was achieved with the IBA threshold set to 0.25. Similar analyses produced the necessary thresholds for high-confidence hints to gene

starts and stops.

Choice of a scoring kernel Computation of the IBA score involves a weighting step, done by a weighting kernel (Equation (3.3)). Figure A.6 compares the accuracy achieved by using (i) the default linear kernel, and (ii) a box kernel.

Maximum number of target proteins per seed gene ProtHint limits the number of target proteins that are splice aligned back to the corresponding seed regions (Section 3.3.2.1). By default, this limit is set to 25; Figure A.7 shows how this parameter affects ProtHint's prediction accuracy.

3.4.3 Comparison of GeneMark-EP+ with GeneMark-EP

The accuracy of GeneMark-EP+ (which enforces the high-confidence hints in the gene prediction step) was about the same as the accuracy of GeneMark-EP when the smallest reference set of proteins—proteins from species outside the phylum of the species in question—was used (Table A.2). The accuracy of GeneMark-EP+ increased significantly when reference proteins from more evolutionarily close species were included; while the accuracy of GeneMark-EP stayed about the same. The only exception was *C. elegans* in which GeneMark-EP's gene-level accuracy dropped by $\sim$4 percentage points for the reference set of species outside the same phylum in comparison with GeneMark-ES (while GeneMark-EP+ showed the accuracy close to the level of GeneMark-ES; Table A.2).

3.4.4 The effect of distinct protein hints on the accuracy of GeneMark-EP+

To differentiate different contributions to GeneMark-EP+'s performance, we compared runs that used only high-confidence intron hints to runs that used only high-confidence hints to gene starts and stops (Table A.5). This experiment showed that enforceable hints of both kinds contributed equally to overall accuracy improvement. However, these hints

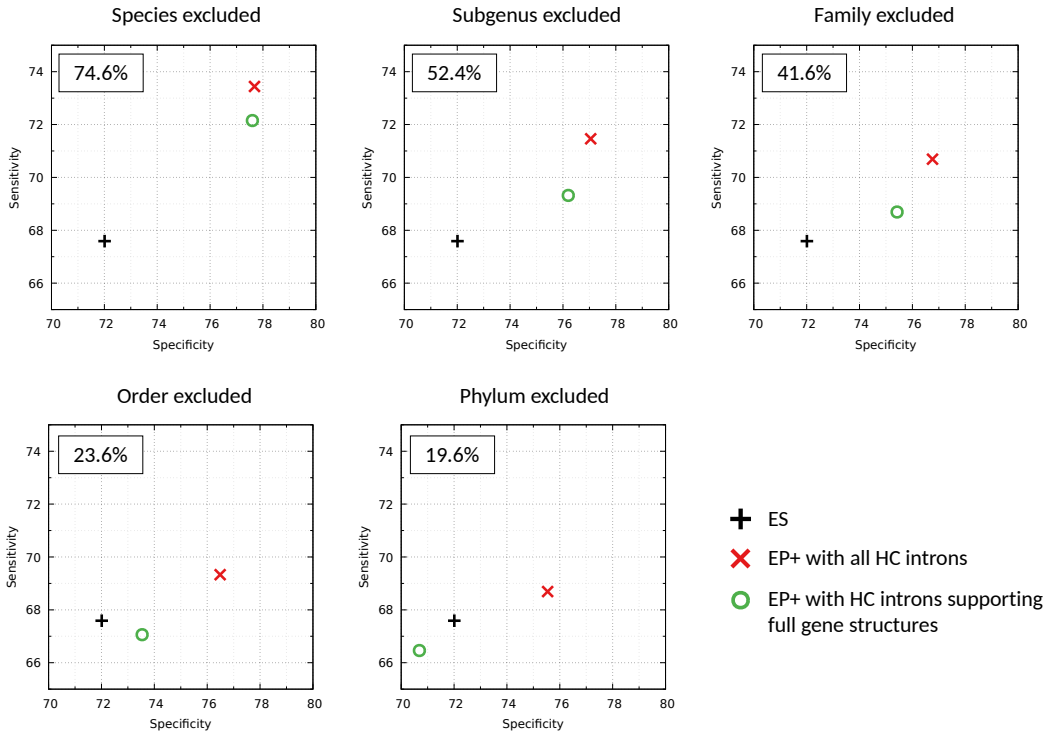

Figure 3.8: A comparison of exon-level accuracy between three gene prediction modes in *D. melanogaster*. The use of introns from incomplete gene alignments led to a significant increase in accuracy compared to using only introns from fully aligned gene structure. GeneMark-ES is represented by a plus symbol. GeneMark-EP+ using only high-confidence (HC) introns is represented by a red cross. GeneMark-EP+ using a subset of HC introns is represented by a green circle. This subset corresponds to annotated gene structures with all the introns supported by HC introns. In each panel, we show the percentage of such introns among all HC introns.

contribute unequally to the reduction of different types of errors. The enforcement of high-confidence intron hints led to a higher prediction accuracy of internal exons; while the enforcement of high-confidence hints to gene starts and stops led to a reduction of errors in initial and terminal exons (Table A.5).

Furthermore, we evaluated the effect of using "partial" vs "complete" protein hints (Section 3.3.4.3). The result of this evaluation (Figure 3.8) shows that GeneMark-EP+ with the full set of protein hints, including the partial ones, consistently outperformed the corresponding runs with complete hints only. This was observed despite the fact that all false positives were removed from the set of complete hints (Section 3.3.4.3). Notably, the accuracy of GeneMark-EP+ with complete hints only was lower than that of GeneMark-ES

	Genes	ES	EP	EP+ Introns (a)	EP+ Starts/Stops (b)	EP+ Full (c)
N. crassa	Merged	129	89	92	64	**74**
	Split	83	96	89	106	**91**
C. elegans	Merged	1120	1076	1090	1019	**1029**
	Split	588	725	614	731	**622**
A. thaliana	Merged	743	634	629	215	**251**
	Split	360	385	242	478	**277**
D. melanogaster	Merged	544	464	462	311	**313**
	Split	285	297	204	324	**221**
S. lycopersicum	Merged	2304	1871	1793	1165	**1192**
	Split	1550	1644	1139	1962	**1252**
D. rerio	Merged	1921	1415	1351	883	**884**
	Split	2553	2976	2018	3058	**2010**

Table 3.6: Numbers of merged and split genes in predictions of GeneMark-ES, -EP and -EP+ with the enforcement of (a) only high confidence hints to introns, (b) only high confidence hints to gene starts and stops (c) enforcement of both (a) and (b). All the numbers were obtained for reference sets of target proteins from the species outside of relevant genus.

when the most evolutionarily remote proteins were used on input (Figure 3.8).

3.4.5 Assessment of gene merging and splitting errors

We observed that GeneMark-ES was more likely to generate gene merging than gene splitting errors (Table 3.6); for instance, a comparison of the *A. thaliana* gene predictions and annotation showed 360 split genes and 743 merged genes. The use of GeneMark-EP (with reference proteins outside the same genus) decreased the frequency of errors in gene merging (a $\sim$15% decrease in all species); however, it also caused a slight increase in gene splitting (Table 3.6). The transition to GeneMark-EP+ (the last column in Table 3.6) reduced gene merging dramatically.

The enforcement of only high-confidence intron hints in GeneMark-EP+ reduced the number of split genes (by enforcing introns in place of incorrectly predicted intergenic regions). Still, these hints had little or no effect on the gene merging (Table 3.6). The most significant effect on gene splitting was observed for *D. rerio*—2010 split genes in the -EP+

mode compared to 2976 in the -EP mode.

The enforcement of high-confidence hints to gene starts and stops significantly reduced the number of merged genes and caused a slight increase in the number of split genes. For instance, the number of merged genes dropped by ∼500 in *A. thaliana* between GeneMark-ES and GeneMark-EP+, a ∼66% improvement; ∼50% improvement was observed for the other species in our tests, except for *C. elegans*. Altogether, in comparison with GeneMark-ES and GeneMarkEP, GeneMark-EP+ achieved a significant reduction in the numbers of both merged and split genes (Table 3.6).

3.4.6 Comparison of GeneMark-EP+ predictions with genome annotations defined by the APPRIS database

We compared GeneMark-EP+ gene predictions with the annotations of major protein isoforms defined by the APPRIS database [85] in genomes of *C. elegans*, *D. melanogaster*, and *D. rerio*. In comparison with our previous assessment using full genome annotations made by respective genomic communities (Table 3.1), this test showed (Figure A.8) an increase in exon-level sensitivity (by ∼4 percentage points for *C. elegans* and *D. rerio*, by ∼7 percentage points for *D. melanogaster*) and a decrease in exon-level specificity (by ∼1.5 percentage points for *C. elegans*, by 3 percentage points for *D. melanogaster* and by ∼8 percentage points for *D. rerio*). The decrease in Sp could be expected since the APPRIS annotation contains a smaller number of exons. The increase in Sn is positive news indicating that GeneMark-EP+, when predicting a single isoform per locus, is likely to predict genes for the major protein isoforms.

At the gene level (Figure A.9), both Sn and Sp were reduced slightly in *C. elegans* and *D. rerio*, and by 5 percentage points in *D. melanogaster*. To correctly interpret this result, we have to remind the definition of gene-level accuracy—a gene is counted as correctly predicted if the prediction matches all exons in at least one annotated alternative transcript. Thus, the removal of alternative transcript isoforms from the reference annotation (as done

in the APPRIS comparison) can only lead to a decrease in gene-level accuracy. The fact that this decrease was small indicates that GeneMark-EP+ mostly predicts the principal gene isoforms.

3.4.7 Comparison of GeneMark-EP/EP+ with -ET

We compared GeneMark-EP/EP+ with GeneMark-ET [48] that uses RNA-Seq short reads to provide external information (hints to intron coordinates) to select anchored gene elements for the GeneMark-ET's algorithm parameter estimation. Notably, GeneMark-ET does not have an "-ET+" mode in which predictions are directly guided by high-confidence hints. We ran GeneMark-ET with hints to coordinates of introns mapped by VARUS [94] from RNA-Seq reads. VARUS automatically sampled, downloaded, and aligned reads from NCBI's Sequence Read Archive (SRA) with the time stamp of 22 January 2020 [93]. The time stamp is important for the reproduction of results since the VARUS outcome depends on the amount of RNA-Seq data deposited to SRA. As one could see (Table A.2), the accuracy of GeneMark-ET with training guided by hints derived from mapped RNA-Seq reads is very close to the accuracy of GeneMark-EP with training guided by hints derived from mapped proteins. The accuracy of GeneMark-EP+, which uses high-confidence hints to improve its predictions, was significantly higher than that of GeneMark-ET in all tests but one (Table A.2); even with the most remote proteins (outside of the same phylum) on input.

3.4.8 More intron hints are generated in regions encoding conserved protein domains

To establish a baseline for this analysis, we first computed how many annotated introns belong to conserved protein domains (using the procedure described in Section 3.3.4.5). We found that $\sim$50% of the whole set of introns annotated in the APPRIS set of principal isoforms are located within conserved protein domains (Table A.6).

In *D. melanogaster*, high-confidence intron hints generated by ProtHint from the species-excluded reference set of proteins fell into regions coding for conserved domains in 55.9%

Table 3.7: For the *D. melanogaster* genome, we show the fractions of high-confidence (HC) intron hints mapped in regions coding for conserved protein domains. The results are provided for sets of reference proteins with different sizes and evolutionary distances to *D. melanogaster*. Out of 41,010 introns in the APPRIS-defined *D. melanogaster* genome annotation, 21,562 (52.6%) are located in regions encoding conserved protein domains.

Exclusion level	High-confidence introns matching APPRIS introns		
	All HC introns	HC introns that fell into domains	
Species	33,894	18,934	(55.9%)
Subgenus	28,437	17,475	(61.5%)
Family	24,670	16,057	(65.1%)
Order	15,829	11,984	(75.7%)
Phylum	9,719	8,222	(84.6%)

of cases (Table 3.7). This fraction increased significantly as more proteins were excluded from the reference set (e.g., proteins from species outside of the *D. melanogaster* genus). Finally, the fraction reached 84.6% when only proteins originating from species outside the *D. melanogaster* phylum were considered (Table 3.7). Similar trends were observed for *C. elegans* and *D. rerio* (Table A.7). In the set of all reported intron hints, the fraction of introns mapped to regions coding for conserved domains was lower than that in the set of high-confidence intron hints (Table A.7); however, the proportion of introns mapped into conserved domain regions also increased upon removing proteins from closely related species.

3.5 Discussion

The main reason to develop GeneMark-EP/EP+ was a clear need to leverage abundant protein sequence data available in public databases for improving the accuracy of automatic gene prediction. It was well expected that the iterative *ab initio* parameterization of statistical models (as done in GeneMark-ES) would become more precise, especially for large genomes, if an efficient method to add data on protein footprints into training was found. This goal was successfully achieved by the GeneMark-EP's semi-supervised train-

ing. Moreover, GeneMark-EP+ significantly improves its prediction accuracy by directly integrating the most confident protein evidence into the predicted exon-intron structures. The need to process large protein databases and determine the reliability of the mapped protein evidence led to the development of a new pipeline called ProtHint. ProtHint finds multiple proteins homologous to a gene initially predicted in a genomic locus and then derives reliable hints to the true gene exon-intron structure by constructing and processing multiple protein footprints.

In this section, we summarize the GeneMark-EP+ results observed in all of the tested species. Next, we highlight the main features contributing to GeneMark-EP+'s high prediction accuracy. We also discuss the main design decisions behind ProtHint that ensure its fast computational speed and high specificity of high-confidence predictions. Finally, we list several limitations of the GeneMark-EP+ approach, most of which are addressed in the remaining chapters of this thesis.

3.5.1 Summary of the GeneMark-EP+ results

The most significant improvement in comparison with GeneMark-ES, observed in all species but the fungus *N. crassa*, occurred when GeneMark-EP+ used the largest possible set of reference proteins (Figures A.2 and 3.7). Although the magnitude of this improvement decreased with the increase of the evolutionary distance of input proteins, the decrease was not dramatic. Even with the most remote proteins, those belonging to species outside of the taxonomic phylum, GeneMark-EP+ was more accurate than GeneMark-ES in all but one case (*C. elegans*).

For *N. crassa*, the use of protein evidence did not lead to a significant difference compared to GeneMark-ES whose high accuracy for fungal genomes was demonstrated earlier [16]. We assume that the relative drop in GeneMark-EP+ performance for *C. elegans* in comparison with *A. thaliana* and *D. melanogaster* was related to a lower number of reference proteins within the *C. elegans* phylum (Table 3.2).

In the genomes of *S. lycopersicum* and *D. rerio*, having longer on average intergenic regions, we observed low exon-level specificity (∼55–60%), likely related to an elevated false positive prediction rate of protein coding genes in the intergenic regions. The gene-level accuracy for *D. rerio*, ∼30% Sn and ∼12% Sp, for any set of reference proteins beyond the *D. rerio* genus, was difficult to improve. Notably, the genes in the *D. rerio* genome have a rather large, 8.2, average number of introns per gene. Under the independence of error assumption, a gene with a large number of introns is a difficult target for an accurate prediction. Even though the independence assumption does not hold in the presence of external evidence, the gene error rate still increases with the increase in the number of introns (data not shown). We also speculate that the relatively lower prediction accuracy observed in the genomes of *S. lycopersicum* and *D. rerio* could be partly attributed to errors in available reference annotations (Section 3.4.1.3). For instance, as shown in Table 3.4, *S. lycopersicum*'s gene-level sensitivity dramatically improved when compared with a more reliable subset of annotation supported by RNA-Seq reads. This was observed despite the fact that RNA-Seq reads were not utilized by GeneMark-EP+.

3.5.2 Sources of accuracy improvements

3.5.2.1 Use of remote homologs

Existing protein-homology-based gene prediction methods, such as GenomeThreader [52], exonerate [51], or GeneWise [54], rely on mapping proteins from *closely related* species to produce predicted exon-intron structures. Thus, their prediction accuracy is dropping fast with the increase of evolutionary distance between the species of interest and the input proteins [54, 57, 63, 64]. In GeneMark-EP+, the simultaneous use of multiple homologous proteins proved to be important for keeping decent accuracy of predictions when the evolutionary distance of input proteins increased. Particularly, due to the corroboration of footprints originating from multiple homologous proteins, we observed an enrichment of high-confidence introns in regions coding for conserved domains (Tables A.7 and 3.7). The

use of partial protein footprints mapped from these domains—when a target protein mapping contributes less than full exon-intron structure—was an important feature of the new method. Partial footprints were useful for expanding the training sets (see Section 3.5.2.2); they also added confident corrections at the gene prediction step (Figure 3.8).

The novel scoring system employed by ProtHint made it possible to define a large number of protein hints with high confidence (>95% Sp; Table A.4), regardless of the evolutionary distance of target proteins. Due to their high specificity, these high-confidence hints could be directly incorporated into the final GeneMark-EP+ predictions; thus significantly increasing the prediction accuracy (Section 3.4.3).

3.5.2.2 Semi-supervised training

Better performance of GeneMark-EP+ in comparison with GeneMark-ES was expected due to two factors: (a) model parameterization on a better-validated training set as the training process becomes semi-supervised instead of unsupervised and (b) enforcement of high-confidence hints in gene prediction steps. Notably, even when direct corrections were not made (GeneMark-EP mode where factor (b) is absent), for all the species but fungi GeneMark-EP showed an improvement over GeneMark-ES (Table A.2).

The use of anchored elements in training was important for the integration of signals originating from different sources (sites predicted from genomic sequence alone and sites identified by protein footprints). The logic of selection of anchored elements enabled filtering of "one-sided" noise present in one or another source. The use of anchored elements was most beneficial for large genomes (*S. lycopersicum* and *D. rerio*; Table A.2) where GeneMark-ES alone generated an elevated rate of false positive errors within long intergenic regions.

Surprisingly, GeneMark-EP showed only small fluctuations in accuracy when the size of the reference set of protein increased by including more evolutionarily close species (Table A.2). This observation suggests that even a relatively small number of anchored

introns play a critical role in parameter estimation and a further increase in the number of anchored introns does not improve the parameters. For the case of *C. elegans*, one could argue that the sufficient minimum number of anchored introns was not found when proteins of the reference set were limited to ones from the species outside the *C. elegans* phylum (Table A.2).

3.5.2.3 Reduction of gene merging errors

The mapping of N- and C-terminals of target proteins allowed for better discrimination between introns and intergenic regions than it could be done by an *ab initio* algorithm. This improvement led to a significant reduction of errors in gene merging (when intergenic regions were incorrectly predicted as introns; Table 3.6). The reduction in error rate of gene splitting (when introns were incorrectly predicted as intergenic regions) was smaller but still significant.

3.5.3 ProtHint design decisions

3.5.3.1 Computational speed

The protein mapping done by ProtHint requires a processing of millions of proteins contained in a protein database. To accelerate this process, we limited the DIAMOND [87] output by 25 target proteins per seed protein as a trade-off between computational speed and prediction sensitivity (Figure A.7).

From the standpoint of runtime reduction, the choices of DIAMOND and Spaln [56] were also critical. DIAMOND is several orders of magnitude faster than BLASTp [97]. The potentially lower alignment precision of DIAMOND was not a concern since ProtHint uses Spaln to generate exact spliced alignment. Nevertheless, we verified that the sensitivity of ProtHint using BLASTp does not significantly differ from the one using DIAMOND. Spaln proved to be the fastest spliced aligner in our tests. A detailed comparison of Spaln and ProSplign (a spliced aligner which can also be used by ProtHint; [53]) is described in

the Appendix in Section A.2.3.1.

3.5.3.2 Ensuring the high specificity of high-confidence hints

GeneMark-EP+ improves over GeneMark-EP (Section 3.4.3) due to the direct influence of enforced hints on prediction steps. The high specificity of high-confidence hints is critical for this improvement to work. Therefore, a significant effort was made to develop the high confidence selection criteria, notably:

- We tested several methods for filtering introns as well as alternative formulas for computing intron borders alignment score (IBA). Longer alignments of individual exons did not produce better intron prediction quality (Figure A.1). The IBA score constructed as an arithmetic mean of upstream and downstream scores S_d and S_u was less accurate than a score using a geometric mean of S_d and S_u (Equation (3.4)).

- IBA and BAQ scores used in the high-confidence hint selection characterize the quality of spliced alignment near the coordinates of a candidate hint. Alignments are weighted by a linear kernel, which gives higher weight to alignment positions close to the coding region boundaries. We tested several other kernels (box, parabolic, triweight); however, the linear one was generating consistently best results for windows of different sizes. A comparison between results of the application of a linear and box kernel is shown in Figure A.6. Window sizes 5, 10, 15, and 20 were tested and 10 was selected as consistently best performing across the species tested.

- The IMC score (mapping coverage) threshold "≥ 4" was tested for proteins from two databases: EggNOG and OrthoDB. In both cases, the cited threshold was leading to similar results across various species tested. The IBA score threshold (0.25) was selected based on the result described in Section 3.4.2.3.

- A comparison between IBA and IMC scores showed that a high value of IMC is a better indicator of high intron specificity than a high value of IBA (Figure 3.2). A

combination of these two scores allowed us to relax the IMC threshold and get a larger set of High-Confidence introns.

- For start codon hints, removing starts overlapped by exons from other protein alignments was critical for ensuring high specificity (Figure 3.5 and Tables A.1 and 3.3).

For even more details about the ProtHint design, see Section A.2 in the Appendix.

3.5.4 Limitations of GeneMark-EP+

3.5.4.1 Accounting for pseudogenes

Since pseudogenes accumulate mutations over years, let us consider groups of "young" and "old". Young pseudogenes, ones with one or two mutations that make them dysfunctional, still have all the sequence patterns that could be used in training. Old pseudogenes, ones that accumulated many mutations, would harm statistical models if included in the training.

We argue that old pseudogenes are rarely predicted by GHMM and the many mutations make them less likely to align against homologous proteins and produce high-scoring protein hints. Therefore, they have little or no chance to be included in the training set of anchored elements (which require both *ab initio* and protein homology support) or in the final predictions.

On the other hand, elements of young pseudogenes could still be identified by GeneMark-ES and mapped by ProtHint. Therefore, the young pseudogenes could positively contribute to parameter training through their "intact" parts. Unfortunately, these "intact" parts would also likely appear in the final predictions. Addressing the full complexity of this issue goes beyond the scope of this project; therefore, currently, GeneMark-EP+ does not collect information on frameshifts and potential pseudogenes.

3.5.4.2 Prediction of alternative isoforms

GeneMark-EP+ searches for a single optimal genomic sequence parse that leads to the prediction of a single gene and a single protein isoform in each locus. The importance of alternative splicing has been debated recently [98], as the evidence was accumulated that a large majority of predicted alternative transcripts may not even be translated into proteins [99]. Moreover, claims were made that when a translated region produces multiple viable protein isoforms, only one among the isoforms, the major one, is expressed in most tissues [86].

If the gene prediction by GeneMark-EP+ is viewed as a prediction of the major isoform, then the result should naturally be assessed in comparison with the annotation of the major isoforms. Such comparison, done for *C. elegans*, *D. melanogaster*, and *D. rerio*, using annotation of principal isoforms provided by the APPRIS database [85], showed improved sensitivity in predicting genes of major protein isoforms; suggesting that GeneMark-EP+ disproportionately predicts the major gene isoforms.

Nonetheless, general tools able to predict all alternative isoforms are of significant interest to the genomic community. Therefore, the task of predicting alternative isoforms is addressed in Chapters 4 and 5.

3.5.4.3 Prediction in genomes with heterogeneous GC content

GeneMark-EP+ does not support multiple models needed for genomes with heterogeneous nucleotide composition, such as genomes of mammals and monocots (e.g., rice and wheat). While the current version of GeneMark-EP+ would perform better than GeneMark-ES when running on such genomes, the overall accuracy could be significantly improved by more accurate modeling of genome heterogeneity. Such modeling is implemented and described in Chapter 5.

3.6 Conclusion

In summary, GeneMark-EP+ should become a universal extension to GeneMark-ES, as its application to a novel eukaryotic genome is facilitated by the use of a vast volume of protein sequences of any evolutionary distance; always available in protein databases such as OrthoDB [66, 67], EggNOG [68], or SwissProt [69]. In the tests on genomes of fungi, plants, and animals, we observed that GeneMark-EP+ delivered better prediction accuracy than *ab initio* GeneMark-ES and RNA-Seq-based GeneMark-ET, even in situations when only evolutionarily remote proteins were used on input.

3.7 Availability

The full GeneMark-EP+ package, including ProtHint, is available at http://topaz.gatech.edu/GeneMark/license_download.cgi. The software is compiled for Linux and Mac OS operating systems. ProtHint is also available as a stand-alone tool at https://github.com/gatech-genemark/ProtHint. All scripts and data used to generate figures and tables in this chapter are available at https://github.com/gatech-genemark/GeneMark-EP-ProtHint-exp.

To give an example, the overall runtime of ProtHint and GeneMark-EP+ on the *D. melanogaster* genome (having $\sim$14,000 genes in a 134 Mbp long sequence) with target proteins selected from species outside the Drosophilidae family was $\sim$5 h on an 8 CPU/8 GB RAM machine. In our experiments, the run-time grew linearly with respect to both genome length and the number of genes.

CHAPTER 4

BRAKER2: AUTOMATIC EUKARYOTIC GENOME ANNOTATION WITH GENEMARK-EP+ AND AUGUSTUS SUPPORTED BY A PROTEIN DATABASE

Abstract

Full automation of gene prediction in eukaryotic genomes has been an important bioinformatics task since the advent of next-generation sequencing. Here we introduce BRAKER2, a protein homology-based gene prediction pipeline integrating ProtHint, GeneMark-EP+, and AUGUSTUS. By combining complementary strengths of these gene prediction tools, BRAKER2 achieves state-of-the-art gene prediction accuracy in a fully unsupervised manner. Under equal conditions, the gene prediction accuracy of BRAKER2 was shown to be higher than the one of MAKER2, one of the most frequently used gene prediction pipelines. Furthermore, in tests with remotely related proteins, the accuracy of BRAKER2 was comparable to that of BRAKER1, which was supported by large amounts of RNA-Seq data.

4.1 Introduction

Constantly improving next generation sequencing (NGS) technology makes it now possible to finish the sequencing of a complete eukaryotic genome within several days. Therefore, accurate automatic methods of genome annotation have been in high demand since the dawn of the NGS era. A self-training algorithm for *ab initio* gene prediction in eukaryotic genomes, GeneMark-ES [16], has accelerated the process of structural annotation for a number of genome projects, e.g., [100–104]. The application of NGS to transcript sequencing (RNA-seq) motivated active development of methods combining genomic and transcriptomic information. A semi-supervised algorithm GeneMark-ET [48] integrated data on spliced aligned RNA-seq reads into GeneMark-ES.

On a parallel avenue, another algorithm, AUGUSTUS [12, 105–108] was demonstrated to be one of the most accurate gene prediction tools [21, 22, 44]. AUGUSTUS carries a flexible mechanism for the integration of external evidence into gene prediction. Furthermore, AUGUSTUS can also use the evidence to predict alternative isoforms. However, like most gene predictors, AUGUSTUS is a supervised algorithm; thus relying heavily on a high-quality training set to estimate its parameters (Section 2.3.1; [1, 16, 17, 19, 47, 61, 62]. The preparation of such training sets requires manual work and validation by experts [13, 16, 17, 47], making it challenging to train AUGUSTUS for novel genomes.

It was apparent that a useful automatic tool could be created by combining strong features of GeneMark-ET and AUGUSTUS. Such a tool, BRAKER1, was developed and released in 2015 [47] to become a frequently used tool in genome annotation projects, e.g., [109–113]. BRAKER1 requires the availability of RNA-seq data; however, not all novel genomes are sequenced along with the species' transcriptomes, e.g., within the Earth BioGenome Project [4]. Moreover, for various reasons, a significant fraction of genes may not be covered by transcripts even if the transcriptome data are generated in the project.

Here, we introduce BRAKER2, a gene prediction pipeline that uses sequences of known cross-species proteins—readily available for any genome project—as external evidence. The use of protein homology in gene prediction poses a challenge due to the patchiness of the evidence proteins generate and the decrease in prediction accuracy with the increasing evolutionary distance of proteins (see Section 2.3.2 for details). Nonetheless, the information contained in large numbers of homologous proteins, including proteins from remotely related species, has the potential to improve genome annotation. GeneMark-EP+, presented in Chapter 3, addressed the challenge by using a large number of cross-species proteins to direct its self-training and gene prediction. To process the input protein database, GeneMark-EP+ uses ProtHint (introduced in Section 3.3.2)—a tool that predicts accurate locations of exon boundaries from a large number of proteins of *any evolutionary distance* to the genome of interest. BRAKER2 integrates complementary strengths of ProtHint,

GeneMark-EP+, and AUGUSTUS to create a fully automated homology-based gene prediction pipeline.

There are several reasons why BRAKER2 performs better than AUGUSTUS or GeneMark-EP+ alone. In contrast to GeneMark-EP+, AUGUSTUS allows for more flexible integration of protein hints into an *ab initio* gene prediction. This makes it possible to integrate *all* of the protein hints generated by ProtHint into AUGUSTUS predictions, not just the subset of high-confidence hints utilized by GeneMark-EP+. On top of that, unlike GeneMark-EP+, AUGUSTUS predicts *alternative isoforms* of protein-coding genes. That said, AUGUSTUS cannot be used at all without a reliable training set—prepared by GeneMark-EP+ in an unsupervised manner. Further, although AUGUSTUS contains a sophisticated mechanism for the integration of protein hints, the task of the actual preparation and scoring of such hints is solved by ProtHint.

In summary, the salient features of BRAKER2 are (i) a fully automatic run, (ii) a massive database search for proteins homologous to proteins encoded in the new genome (yet unknown ones), (iii) processing of millions of protein to genome spliced alignments to generate hints to exon-intron structures, and (iv) integration of genomic sequence patterns and protein hints to the gene structure at all iterative steps of model training and gene prediction.

We assessed the prediction accuracy of BRAKER2 on well-studied and, arguably, well-annotated genomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. For additional tests on nine additional genomes, we selected subsets of annotated genes corroborated by RNA-Seq evidence. Further, we compared the performance and accuracy of BRAKER2 to that of MAKER2 [114], the most frequently used competing gene prediction pipeline, as well as to BRAKER1.

4.2 Materials

For testing BRAKER2, we used genomic sequences and gene annotations of twelve species. Among them were the early sequenced model organisms: *A. thaliana*, *C. elegans* and *D.*

Table 4.1: Genomes used in the tests; asterisks indicate model organisms. An average number of introns per gene was determined with respect to the number of all the annotated genes in the genome. For a gene to be considered complete and canonical, at least one of the gene's transcripts had to be fully annotated, such that the initial coding exon started with a "canonical" ATG and the terminal coding exon ended with TAA, TAG, or TGA.

Species	Annotation version	Genome size (Mb)	# Genes in annotation	# Introns per gene	% Non-canonical or incomplete genes
Species with early sequenced genomes					
*A. thaliana**	Tair Araport 11 (Jun 2016)	119	27,445	4.9	0.3
*C. elegans**	WormBase WS271 (May 2019)	100	20,172	5.7	0.2
*D. melanogaster**	FlyBase R6.18 (Jun 2019)	138	13,929	4.3	0.3
Other species					
Plantae					
*P. trichocarpa**	JGI Ptrichocarpa_533_v4.1 (Nov 2019)	389	34,488	4.9	0.3
*M. truncatula**	MtrunA17r5.0-ANR-EGN-r1.6 (Feb 2019)	430	44,464	2.9	0.0
S. lycopersicum	Consortium ITAG4.0 (May 2019)	773	33,562	3.5	14.5
Arthropoda					
B. terrestris	NCBI Annotation Release 102 (Apr 2017)	249	10,581	7.1	4.7
R. prolixus	VectorBase RproC3.3 (Oct 2017)	707	15,061	4.8	34.7
P. tepidariorum	NCBI Annotation Release 101 (May 2017)	1,445	18,602	7.3	18.2
Vertebrata					
T. nigroviridis	TETRAODON8.99 (Nov 2019)	359	19,589	10.4	63.8
*D. rerio**	Ensembl GRCz11.96 (May 2019)	1,345	25,254	8.2	11.8
*X. tropicalis**	NCBI Annotation Release 104 (Apr 2019)	1,449	21,821	12.1	2.4

melanogaster. The other nine species were: the plants *Populus trichocarpa*, *Medicago truncatula*, *Solanum lycopersicum*, the arthropods *Bombus terrestris*, *Rhodnius prolixus*, *Parasteatoda tepidariorum*, and the vertebrates *Tetraodon nigroviridis*, *Danio rerio*, *Xenopus tropicalis* (Tables B.1 and 4.1). In all genomic datasets, contigs not assigned to any chromosome and the genomes of organelles were excluded from the analysis.

RNA-seq data used in the accuracy evaluation and runs of BRAKER1 were sampled from the Sequence Read Archive [93] by VARUS [94]. To determine to which degree both predicted and annotated genes covered the sets of universal single-copy genes identified by the BUSCO protein families, we used the BUSCO database v4 [115].

4.2.1 Protein database preparation

We used the OrthoDB database v10 [66] as a source of protein data. A test of BRAKER2 on a well-studied genome should utilize a set of cross-species proteins that imitates a protein

Table 4.2: Composition of the clades of OrthoDB v10 used by BRAKER2. Numbers in black bold show the largest numbers of species used to support gene predictions for a given species (left column). The numbers of species removed from the largest OrthoDB segment in evaluation assessments are shown in blue. Species whose proteins are not present in OrthoDB v10 are marked with asterisks.

Species	# of species in the OrthoDB clade						Name of the largest OrthoDB segment	# of proteins in the OrthoDB segment
	Genus	Family	Order	Class	Phylum	Kingdom		
A. thaliana	**2**	**8**	**10**	-	100	**117**	Plantae	3,510,742
C. elegans	**3**	**3**	**5**	6	7	**448**	Metazoa	8,266,016
D. melanogaster	**20**	**20**	**56**	148	**170**	**-**	Arthropoda	2,601,995
*P. trichocarpa**	1	5	**5**	-	100	**117**	Plantae	3,510,742
M. truncatula	1	10	**10**	-	100	**117**	Plantae	3,510,742
S. lycopersicum	2	10	**11**	-	100	**117**	Plantae	3,510,742
*B. terrestris**	1	7	**40**	148	**170**	-	Arthropoda	2,601,995
R. prolixus	1	1	**16**	148	**170**	-	Arthropoda	2,601,995
P. tepidariorum	1	1	**2**	10	**170**	-	Arthropoda	2,601,995
*T. nigroviridis**	0	1	**1**	50	**246**	-	Chordata	5,003,104
D. rerio	1	5	**5**	50	**246**	-	Chordata	5,003,104
X. tropicalis	2	2	**3**	3	**246**	-	Chordata	5,003,104

set available for running BRAKER2 on a newly sequenced genome. Proteins that originate from the most evolutionarily close species are expected to be most informative for the BRAKER2 algorithm. Therefore, a meaningful characteristic of a selected set of reference proteins is the least evolutionary distance of proteins from the reference genomes to the genome in the test.

To make these selections for *A. thaliana*, *C. elegans*, and *D. melanogaster*, we started from large clades (Plantae, Metazoa and Arthropoda, respectively) and created three sets of proteins for each species by excluding either (i) proteins from the given species per se, (ii) proteins from all species of the same family, (iii) proteins from all species of the same order. For the other nine species, we also defined the large clades and then only used partitions of type (iii) (Table 4.2).

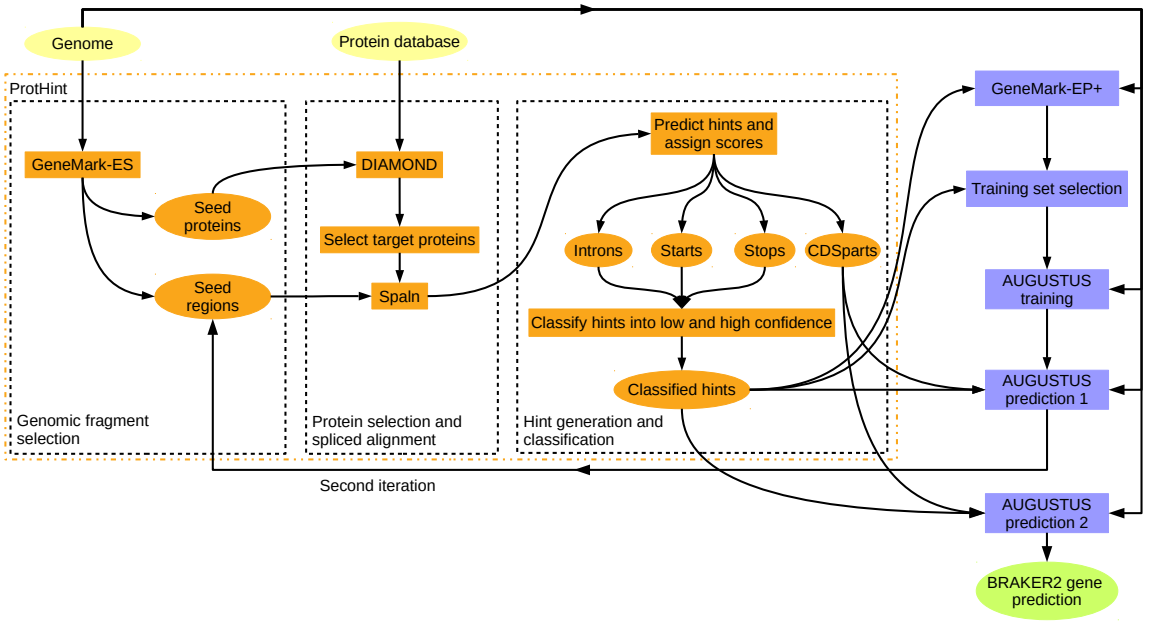

Figure 4.1: Flowchart of the BRAKER2 pipeline. Input, intermediate and output data are shown by ovals. The tools and processes of the ProtHint pipeline are shown in orange; other components of BRAKER2 are shown in blue.

4.3 Methods

4.3.1 Overview of BRAKER2

The overview of BRAKER2 is shown in Figure 4.1. At the first step, BRAKER2 executes the ProtHint protein mapping pipeline (described in detail in Section 3.3.2). In BRAKER2, in addition to the original hint generation scheme, ProtHint makes hints called *CDSpart chains*. This hint type helps to correctly combine exons predicted by AUGUSTUS into a single transcript (details in Section 4.3.3). The hints generated by ProtHint are used by GeneMark-EP+ to execute its self-training (Section 3.3.3.1) and gene prediction steps (Section 3.3.3.2). From the whole complement of genes predicted by GeneMark-EP+, BRAKER2 selects a set of anchored genes that contain GeneMark-EP+ predictions supported by protein hints (Figure 4.2; Section 4.3.2). The anchored genes are used to train AUGUSTUS. Once trained, AUGUSTUS is used to predict genes in the genomic DNA. At the stage of gene prediction, AUGUSTUS enforces all high-confidence hints defined by Prot-

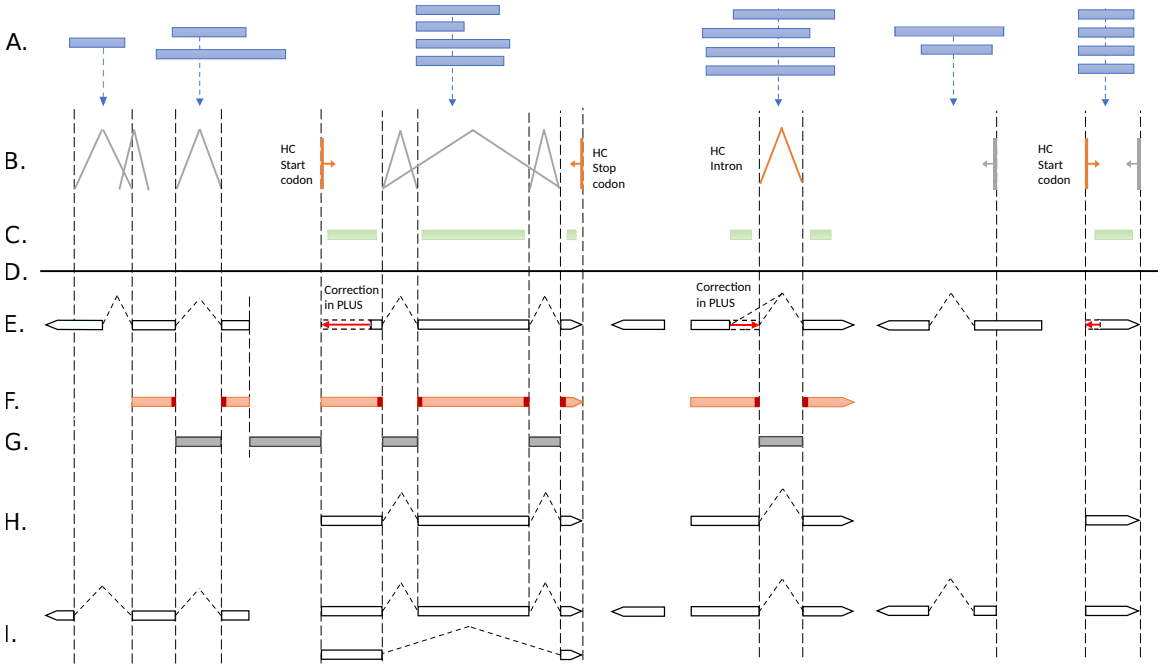

Figure 4.2: Evidence integration in BRAKER2. (A) Target proteins; (B) Introns, gene start and stop sites defined by spliced alignments of target proteins to genome; (C) CDSpart chains; (D) Genome sequence; (E) Genes predicted by GeneMark-EP+ at a given iteration. The high-confidence hints are enforced (red arrows); (F) Anchored sites—the splice sites and gene ends predicted *ab initio* and corroborated by protein hints; (G) Anchored introns and intergenic sequences bounded by anchored gene ends, selected for the training of a non-coding sequence model of GeneMark-EP+; (H) GeneMark-EP+ multi-exon and single-exon genes anchored by protein hints—selected for the training of AUGUSTUS; (I) Transcripts predicted by AUGUSTUS with the support of protein evidence.

Hint (Section 3.3.2.3). Furthermore, the CDSpart chain and non-high-confidence hints are integrated into the AUGUSTUS gene prediction as well (Figure 4.1; Section 4.3.3). In the genomic regions lacking hints from cross-species protein alignments, GeneMark-EP+ and AUGUSTUS predict genes in an *ab initio* mode. At the end, BRAKER2 executes a second major iteration in which ProtHint and AUGUSTUS are re-run using information from the first iteration. The genes predicted by AUGUSTUS in the second iteration constitute the final prediction set of BRAKER2.

4.3.2 The selection of AUGUSTUS training sets

Genes predicted by GeneMark-EP+ are filtered and sampled prior to training AUGUSTUS in the following way:

1. The ratio of multi-exon and single-exon genes is determined prior to filtering.

2. During filtering, multi-exon genes are retained if they are anchored, i.e., have support by an intron hint from at least one protein alignment for every intron (Figure 4.2).

3. The minimum number of required single-exon genes in relation to filtered multi-exon genes is computed to keep the proportion from step 1 in step 4.

4. Single-exon genes are selected if they are anchored, i.e., have support from protein evidence in terms of start- and stop-codon hints (Figure 4.2). If the number of the selected single-exon genes is lower than the minimum required number of single-exon genes (step 3), additional single-exon genes predicted by GeneMark-EP+ lacking protein evidence support are randomly added to reach the minimum number.

5. If the resulting number of training genes is lower than 4000, additional genes are added in the diminishing order of their support rank by protein hints. A gene support rank is computed as follows:

$$S_r = \frac{\#of\ supported\ borders\ of\ protein\text{-}coding\ exons}{\#of\ actual\ borders\ of\ protein\text{-}coding\ exons} \tag{4.1}$$

   Genes are then added in the descending order of their S_r.

6. Complex genes with many introns contribute more effectively to the training of AUGUSTUS than gene structures with few or no introns. Thus, "simple" genes with few or no introns are down-sampled as described in [116].

7. Training genes are translated into protein sequences that are searched against each

other. If two sequences have an identity of more than 80%, one gene is removed from the training gene set. The similarity search is executed using DIAMOND [87].

8. If, at the end, there are more than 8000 training gene structures, genes are randomly down-sampled to 8000 genes. This is done to decrease the runtime.

Thus prepared training gene set is then randomly split into three sets:

1. A set for running *etraining*, a tool for training AUGUSTUS parameters,
2. a set for evaluating AUGUSTUS meta-parameter optimization steps, and
3. a test set. This set is used as an independent test set for estimating the accuracy.

If the total number of genes is smaller than 600, $1/3$ of all available genes are sampled into each set. If there are 600 to 1000 available gene structures, 200 genes each are sampled into the last two sets, all remaining genes go into the first set. If there are more than 1000 training gene structures, 300 genes each are sampled into the last two sets, all remaining genes go into the first set. AUGUSTUS is then trained on these sets as described in [116].

4.3.3 Integration of protein hints

The types of protein hints generated by ProtHint and utilized by BRAKER2 in the final gene prediction step of AUGUSTUS are shown in Figure 4.3. This section describes unique aspects of hints integration that are implemented in BRAKER2 but were not present in GeneMark-EP+.

4.3.3.1 Chained hints

GeneMark-EP+ and AUGUSTUS treat intron as well as start and stop hints as independent ones, not necessarily related to one and the same gene. On the other hand, the CDSpart hints, which specify the locations of protein-coding exons, are treated as a "chain" of evidence when originating from the same protein. AUGUSTUS attempts to incorporate

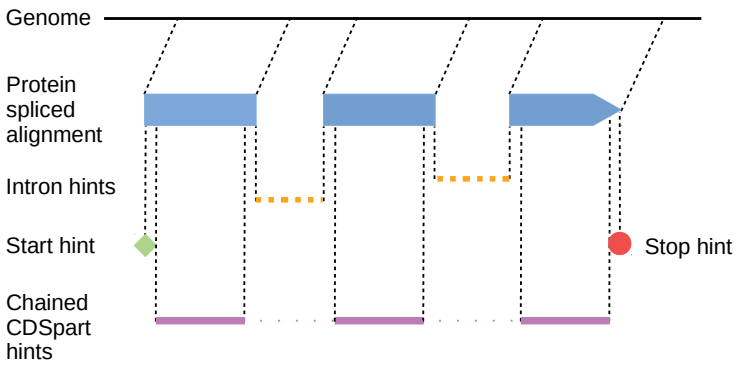

Figure 4.3: Schematics of the types of hints used in BRAKER2 (introns, start and stop, CDSpart) derived by ProtHint from a spliced alignment of a protein to genomic sequence.

CDSparts from the same chain into the same transcript model. Thus, correctly chained hints help to reduce gene splitting errors and aid in the correct identification of alternative isoforms.

ProtHint prepares a chain of exon hints for each seed gene based on the spliced alignment of the highest-scoring target protein. In this process, all exons with an AEE score < 25 are removed from the chain (see Section 3.3.2.2 for the definition of the AEE score). Further, each exon hint is trimmed at its boundaries by 15 nucleotides (hence the "part" in *CDSpart* hints). The trimming is done because, in ProtHint, the exact exon-intron boundaries are better captured and scored by hints to introns and start/stop codons.

We experimented with using more than one chain per seed gene; however, this only led to a significant increase in the runtime of AUGUSTUS without increasing the overall prediction accuracy.

4.3.3.2 Integration of all protein hints

Apart from the high-confidence ProtHint hints, which are enforced in the prediction steps of both GeneMark-EP+ and AUGUSTUS, AUGUSTUS also utilizes all the remaining protein hints to increase the likelihood of gene structure candidates that are in agreement with the hints. BRAKER2 further separates the non-high-confidence hints into two categories:

medium- and low-confidence (with medium-confidence hints having a bigger effect on the likelihoods of AUGUSTUS predictions). For this separation, BRAKER2 uses logistic regression with parameters that were obtained with ProtHint using the *D. melanogaster* genome and Arthropoda section of OrthoDB.

Specifically, to train the logistic regression parameters, ProtHint intron hints were labeled as true or false using the reference annotation. A hold-out test set of 500 hints was set aside. Hints were predicted as true or false based on their IMC and IBA scores (Section 3.3.2.2) by fitting a logistic regression model. The accuracy of the model was checked on the hold-out test set and determined to be 93% (the proportion of true positives in the test set was 80%). We also verified that the classification procedure worked well across the other tested species. The logistic regression training procedure for start and stop codon hints was analogous, using the SMC and BAQ scores (Section 3.3.2.4).

The non-high-confidence hints classified as true by the logistic regression model are put into the medium-confidence set, while the rest goes into the low-confidence set. The influence hints in both sets have on modifying the likelihood of AUGUSTUS predictions (i.e., the hints weights) was determined by supervised training in multiple genomes. The final hint weights, used by BRAKER2 in all runs, are shown in the Appendix in Section B.1.

4.3.4 Second iteration of BRAKER2

BRAKER2 runs in two major iterations (Figure 4.1). The first one starts with ProtHint using seed genes predicted by GeneMark-ES [16]. Seeds for some true genes might be missed at this stage; therefore, to recover potentially lost seeds, BRAKER2 runs a second iteration that uses the genes predicted in the first iteration as seed genes. In the second iteration, ProtHint runs the database search and other processing only for seed genes that differ from the original seeds and merges the newly defined hints with hints from the first iteration. Then, AUGUSTUS uses the models trained in the first iteration along with the updated protein hints to predict the final set of genes. The second BRAKER2 iteration has

fewer steps (no GeneMark-EP+ or AUGUSTUS training, ProtHint is run only for unique seeds); consequently, it does not significantly increase BRAKER2's overall runtime.

4.3.5 Methods related to algorithm assessment

This section describes the design of methods that were used to evaluate aspects of BRAKER2; other than the standard accuracy assessment described in Section 2.4. In all evaluations, regions of annotated pseudogenes were excluded from comparisons.

4.3.5.1 Repeat masking

All tested gene prediction algorithms (including BRAKER2 itself) were run on repeat-masked genomes. Repetitive sequences (interspersed repeats and low complexity sequences) were identified by RepeatModeler [91] and RepeatMasker [92]. A run of RepeatModeler on a whole genome produced a repeat library. Next, the locations of repeats were identified and soft-masked by RepeatMasker.

Repeat masking by RepeatModeler and RepeatMasker with default settings was sufficient to achieve high gene prediction accuracy in all the tested genomes except for *X. tropicalis*. *X. tropicalis* contained a large number of long tandem repeats ($\sim$60 Mb in total) with elevated GC content (Figure B.2) that were not identified by the default RepeatModeler/RepeatMasker run. When left unmasked, these repeats caused GeneMark-ES (running in the initial step of BRAKER2) to converge to an incorrect statistical model of a protein-coding region and to make incorrect gene predictions. Particularly, GeneMark-ES would predict a majority of coding exons (93%) in the GC-rich regions of the long tandem repeats and would poorly predict the true *X. tropicalis* genes. To solve this problem, we identified and masked the problematic *X. tropicalis* repeats by an additional run of Tandem repeats finder [117] with the maximum repeat period size set to 500.

4.3.5.2 *Selection of reliable annotation subsets*

To assess the accuracy of gene prediction made for *A. thaliana*, *C. elegans*, and *D. melanogaster*, we compared the predicted genes with the full complement of the reference annotation. For the nine additional genomes (Table 4.1), which arguably have less reliable reference annotations [3], we prepared more reliable annotation subsets by selecting genes that (i) were complete, and (ii) had all their introns supported by RNA-Seq hints mapped by VARUS. The prediction sensitivity was then evaluated with respect to both the reliable subset and the full reference annotation.

4.3.5.3 *Use of universal single-copy genes from BUSCO families*

The BUSCO metric evaluates the completeness of a genome assembly and annotation; it is based on collections of single-copy genes expected to be present in a particular lineage [118]. The "BUSCO genes" may constitute $< 5\%$ of genes in a genome; nonetheless, this approach is practical for novel genomes given its relatively easy application. We used the BUSCO metric to characterize gene sets predicted by BRAKER2 in the nine genomes with less reliable annotations.

When using BUSCO, it is important to understand the limitations of BUSCO as a gene prediction accuracy evaluation tool. While the BUSCO metric gives an estimate of the gene prediction algorithm's Sn value, it does not quantify an algorithm's tendency to predict false positives (the Sp value). Moreover, since BUSCO relies on an HMMER3 [119] search for detecting homologs of the BUSCO proteins, it does not discriminate between precisely and approximately predicted exon-intron structures. Therefore, the BUSCO metric is less precise in the assessment of gene prediction accuracy than the methods comparing the coordinates of predicted and annotated genes.

4.3.5.4 *Assessment of the AUGUSTUS training set selection*

We conducted several experiments to assess the effect of selecting various AUGUSTUS training sets on its prediction accuracy. First, we compared the accuracy of AUGUSTUS trained on all vs anchored GeneMark-EP+ genes (Section 4.3.2). We further evaluated the effect of (i) using training genes directly from the reference annotation, and (ii) using only highly conserved genes. For this experiment, we used GeneMark-EP+ predictions made in *D. melanogaster*, having only remote proteins on input (proteins from outside of the taxonomic phylum). From the set of GeneMark-EP+ genes, we prepared four distinct training sets: (i) anchored genes (the default training set), (ii) randomly sampled GeneMark-EP+ genes, (iii) randomly sampled true positive genes (as determined by comparison with the reference annotation), and (iv) highly conserved genes—genes that had all introns supported by high-confidence hints mapped by ProtHint. All the sets contained the same number of training genes (achieved by random downsampling). Finally, we evaluated the effect of the number of anchored genes on the AUGUSTUS training by using 500, 1000, 2000, . . . , 10,000 anchored training genes (randomly sampled from the full set of anchored genes). This experiment was done using the genome of *A. thaliana* since BRAKER2 generated more than 10,000 anchored genes for this genome.

4.3.5.5 *The effect of increasing the number of species in the reference protein database*

To demonstrate that the increase of the number of species in the reference protein set is a positive factor for the accuracy of predictions, we conducted the following two experiments. In the first experiment, we first sorted the species in the Drosophila genus by their average protein similarity to *D. melanogaster*. Next, we ran BRAKER2 five times and evaluated the prediction accuracy of each run. In the first run, only proteins from the 5th most taxonomically distant Drosophila, *D. virilis*, were used. Next, we added proteins from a more remote *D. mojavensis* to the set. In the third experiment, we included *D. hydei*, etc.

In the second experiment, we ran BRAKER2 on *D. melanogaster* with three different

input protein sets: (i) proteins from the Anopheles genus (14 distinct species), (ii) all proteins of species outside of *D. melanogaster*'s taxonomic family, and (iii) proteins of species outside of *D. melanogaster*'s taxonomic order. Importantly, Anopheles species are outside of the *D. melanogaster* taxonomic family but in the same taxonomic order (Figure B.7). The main goal of this experiment was to see whether using the larger number of proteins in set (iii) can compensate for the benefit of having closer relatives in set (i).

In both experiments, the similarity between *D. melanogaster* and other species was estimated as follows. We first aligned proteins of *D. melanogaster* against all target Arthropoda proteins in OrthoDB (which included proteins of *D. melanogaster* itself) with DIAMOND. Next, for each target species s, we computed the sum of bitscores Bit_s of all alignments involving a target protein of s. The similarity measure $Sim(s, melanogaster)$ between a species s and *D. melanogaster* ($melanogaster$) was then determined as:

$$Sim(s, melanogaster) = \frac{Bit_s}{Bit_{melanogaster}} \tag{4.2}$$

4.3.5.6 Predicting genes with BRAKER1

BRAKER1 is a genome annotation pipeline that combines self-training GeneMark-ET with AUGUSTUS [47]. BRAKER1 uses external evidence in the form of introns originating from short RNA-Seq reads mapped to genome. BRAKER1 was run on the genomes of *A. thaliana*, *C. elegans* and *D. melanogaster*. The hints for BRAKER1 were prepared by VARUS [94]; the details of the VARUS runs are described in Section B.2.

4.3.5.7 Predicting genes with MAKER2

The MAKER2 genome annotation pipeline can combine information from several sources, such as *ab initio* gene predictions, mapped RNA-Seq reads, as well as alignments of proteins to the genome [114, 120, 121].

For our tests, we chose genomes of *A. thaliana*, *C. elegans*, and *D. melanogaster*, ar-

guably the best-annotated genomes among the ones we considered. For each genome, we used the relevant segment of the OrthoDB database as described in Section 4.2.1, with the exclusion of species of the same taxonomic order. All the components of the MAKER2 pipeline (e.g., the repeat annotation or training of gene finders), were executed in a *de novo* mode; i.e., each of the three genomes was considered to be a "novel" one.

Because of MAKER2's design, the protein mapping in MAKER2 is much slower than protein mapping done by ProtHint in BRAKER2; therefore, we further limited each of the three OrthoDB partitions to randomly selected ten species (Table B.2). To make a fair comparison, all comparisons between MAKER2 and BRAKER2 were done with BRAKER2 using the same limited protein database.

We used two MAKER2 execution protocols (Figure B.2). In the first protocol (Figure B.2A), recommended by the authors [121], the protein spliced alignments were used to create training sets for AUGUSTUS and SNAP [13]. The final gene predictions were made by combining predictions of self-training GeneMark-ES with the ones of AUGUSTUS and SNAP, both using the protein-derived hints. Based on the experience with developing BRAKER2, we introduced a second training protocol (Figure B.2B). In this protocol, which we called BRAKER2-like, protein spliced alignments and GeneMark-ES predictions were used to create a training set for AUGUSTUS. The final gene predictions were done by only two gene finders: GeneMark-ES and AUGUSTUS with hints.

MAKER2 offers two modes of gene prediction: to only get predictions supported by external evidence or to report all predictions, including *ab initio* predictions generated without support. We executed MAKER2 in the second mode, the one producing higher Sn values. This protocol is also more similar to BRAKER2, which reports all predictions.

The repeat masking for both BRAKER2 and MAKER2 was done with the same genome-specific repeat libraries (generated by RepeatModeler; Section 4.3.5.1). Both BRAKER2 and MAKER2 training and predictions were done on repeat-masked sequences. However, it is important to note that even though BRAKER2 and MAKER2 used the same repeat

libraries, the two tools have different methods of processing repeat masking.

For the BRAKER2 runs, the sequences were soft-masked with RepeatMasker (Section 4.3.5.1). Within BRAKER2, AUGUSTUS uses information on soft-masked regions at the gene prediction step and reduces the probability of coding exon predictions in repeat regions. GeneMark-ES and -EP+ do hard-masking of soft-masked repeats longer than 1000 nt (or $>$ 100 nt for genomes $>$ 300 Mbp). The topic of repeat-masking in GeneMark is addressed in detail in Chapter 5.

MAKER2 runs RepeatMasker internally. Subsequently, MAKER2 does hard-masking of all interspersed (complex) repeats, while low-complexity (simple) repeats remain soft-masked. Borders of complex repeats are extended by 50 nt. To evaluate the effect of MAKER2's rather strict hard-masking on its prediction accuracy, we also executed the prediction steps of MAKER2 on unmasked genomes.

4.4 Results

We evaluated BRAKER2 on the twelve genomes described in Section 4.2, using the accuracy evaluation methods presented in Section 4.3.5. Importantly, we executed BRAKER2 with different protein sets on input, to simulate the absence of closely related species in the input protein database (see Section 4.2.1).

4.4.1 Accuracy assessment of BRAKER2 and comparison with BRAKER1

4.4.1.1 Genomes of A. thaliana, C. elegans, and D. melanogaster

The accuracy of BRAKER2 was determined at gene and exon levels (Figures 4.4 and 4.5 and Tables B.3 to B.5). The gene-level Sn and Sp were determined in comparison with the reference annotations of *A. thaliana*, *C. elegans*, and *D. melanogaster*, and showed the following patterns (Figure 4.4). BRAKER2 always performed better than BRAKER1 (by $\sim$12 percentage points in terms of average F1 gene-level accuracy) when BRAKER2 used the largest set of reference proteins—excluding only proteins of the same species (Tables B.3

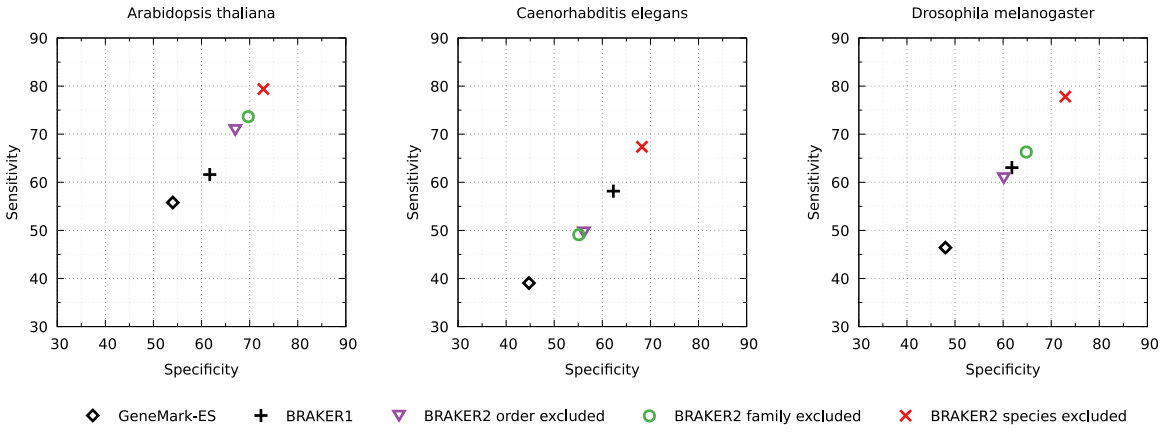

Figure 4.4: Gene-level Sn and Sp, corresponding to three runs of BRAKER2 with protein support, a run of BRAKER1 with RNA-seq support, and a run of GeneMark-ES. BRAKER2 was run with the support of proteins from OrthoDB excluding proteins (i) of the same species, (ii) of all species of the same taxonomic family, and (iii) of all species of the same taxonomic order.

to B.5). With the smaller protein sets, those excluding proteins from all species of the same family or of the same order, the comparison between BRAKER1 and BRAKER2 was more mixed. On *A. thaliana*, BRAKER2 always outperformed BRAKER1, irrespective of the input protein set used by BRAKER2. On *D. melanogaster*, BRAKER2 performed better than BRAKER1 when proteins of the same taxonomic family were excluded from the protein database, but slightly worse when proteins from the same order were excluded. Finally, on *C. elegans*, BRAKER2 only outperformed BRAKER1 when the proteins of the same species were excluded. The patterns of accuracy change on the gene level mainly translated into the patterns observed at the exon level (Figure 4.5 and Tables B.3 to B.5). The numbers of alternative isoforms predicted by both BRAKER1 and BRAKER2 are shown in Table B.7.

4.4.1.2 Additional set of test genomes

Model organisms *A. thaliana*, *C. elegans*, and *D. melanogaster* were subjects of the pilot genome sequencing projects; therefore, we used their longtime curated genome annotations as whole genome test sets. In conducting tests on genomes of the other nine species

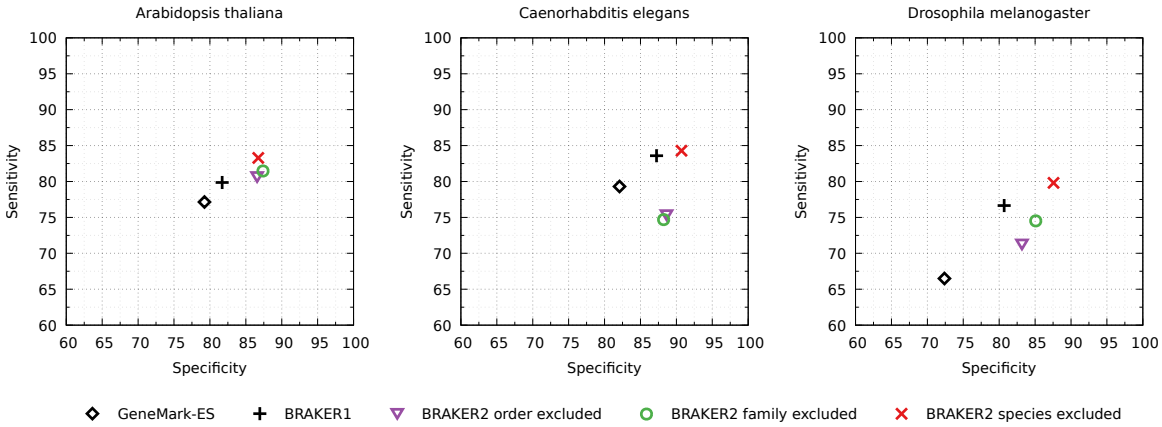

Figure 4.5: Exon-level Sn and Sp for the same tests as shown in Figure 4.4.

Table 4.3: Gene prediction sensitivity of BRAKER2 at the gene and exon levels. The test sets were: (All) all annotated multi-exon genes, and (Reliable) all annotated complete multi-exon genes having all introns supported by mapped RNA-seq reads.

Species	Gene Sn		Exon Sn		% Reliable Genes
	All	Reliable	All	Reliable	
A. thaliana	70.2	78.8	81.5	87.9	83.5
C. elegans	49.8	57.8	75.7	81.0	81.1
D. melanogaster	59.5	61.6	71.9	74.4	93.2
P. trichocarpa	69.3	76.4	86.2	90.4	84.6
M. truncatula	48.3	63.2	82.7	90.0	69.6
S. lycopersicum	40.7	68.0	78.5	92.1	54.4
B. terrestris	45.7	56.7	74.6	79.5	75.1
R. prolixus	13.2	45.5	61.4	80.2	26.4
P. tepidariorum	24.6	40.2	67.9	79.9	50.6
T. nigroviridis	10.4	67.7	60.6	89.5	11.2
D. rerio	39.1	50.3	75.6	86.3	70.8
X. tropicalis	38.9	46.3	75.3	80.0	74.8

(the blue-colored names in Table 4.3), we used a different approach (see Section 4.3.5.2) motivated by the following example. Upon comparison of the gene predictions made by BRAKER2 in the *R. prolixus* genome with its reference annotation (Table 4.1), the gene-level Sn value appeared to be 13.2% (Table 4.3). However, the Sn value was 45.5% when computed against a set of multi-exon *R. prolixus* genes with all introns supported by at least one mapped RNA-seq read (a 26.4%-large subset of all multi-exon genes).

In seven out of the nine additional genomes (the exceptions being *P. trichocarpa* and *X. tropicalis*), large improvements ($>$ 10 percentage points) in the gene-level Sn values were

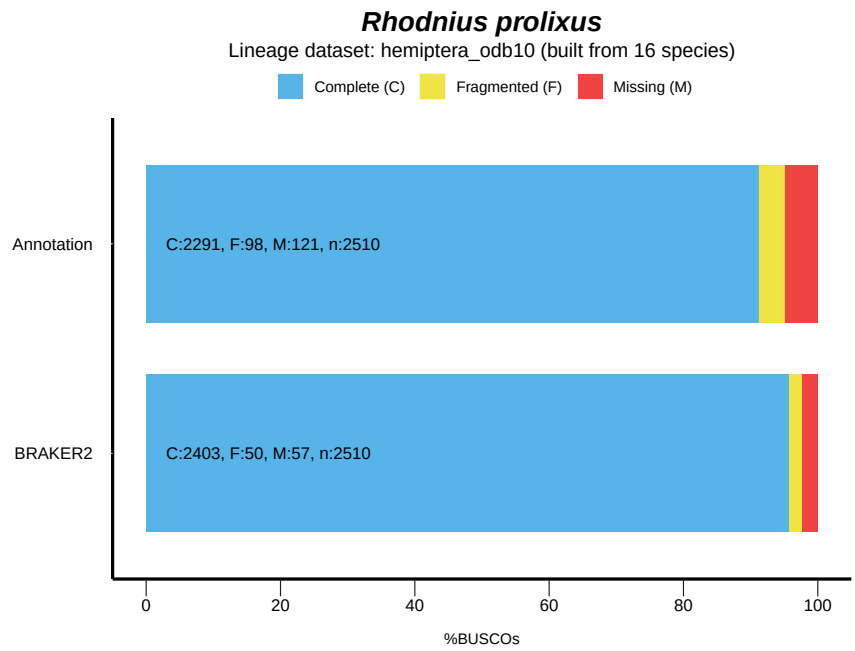

Figure 4.6: Statistics of the sets of genes from BUSCO families (complete, fragmented, missing) identified in the reference genome annotation of *R. prolixus* (top); the same statistics for the set of genes predicted by BRAKER2 (bottom).

observed when the base for comparison was changed from a whole set of annotated genes to the narrower (reliable) set (Table 4.3). Reassuringly, such an effect was not observed for *A. thaliana*, *C. elegans*, and *D. melanogaster*. Overall, in the test sets of genes supported by the mapped RNA-seq reads, we observed exon Sn values near 80% for the three arthropods, between 80 and 87% for the three vertebrates, and the highest, close to 90%, for the three plants (Table 4.3). Detailed comparisons against full complements of the reference annotations are shown in Table B.6.

Among the genes predicted by BRAKER2 in each of the nine genomes, we also identified genes encoding proteins from the species-specific BUSCO protein families (Section 4.3.5.3). For a given genome, a percentage of such recognized "BUSCO members" among the full species-specific BUSCO set provided an estimate of the sensitivity of a gene prediction method (assuming no errors in assembly) and was be compared with the corresponding figures determined for the reference genome annotation (Figures B.3 to B.5 and 4.6).

In the plant and arthropod genomes, BRAKER2 missed ∼3% or less of the BUSCO genes (Figures B.3 and B.4). Moreover, fewer BUSCO genes were missed by BRAKER2 than by the current reference annotations of genomes of *M. truncatula*, *S. lycopersicum*, *P. tepidariorum*, and *R. prolixus* (Figure 4.6). The percentages of BUSCO genes missed by BRAKER2 in the vertebrate genomes were: ∼12% in *T. nigroviridis*, ∼5% in *D. rerio*, and ∼9% in *X. tropicalis* while the genome annotations missed ∼12, 3, and 3%, respectively (Figure B.5).

4.4.2 Effect of the selection of training genes on gene prediction accuracy

We compared the effect of using all vs anchored GeneMark-EP+ training genes on AUGUSTUS accuracy (see Section 4.3.5.4). The use of anchored sets improved the gene level F1 values of *ab initio* gene prediction by AUGUSTUS in *A. thaliana*, *C. elegans*, and *D. melanogaster* genomes by two to five percentage points (Table 4.4). We cite here the accuracy of *ab initio* gene prediction since the full BRAKER2 could get further improvements from the external protein hints, which would overshadow the effects of training. The use of anchored genes for the AUGUSTUS training had an even stronger effect for the large genomes where the difference in F1 value at both exon and gene levels for *D. rerio* reached ∼10 percentage points (Table 4.4).

We further evaluated the effect of (i) using training genes from the reference annotation, and (ii) using only highly conserved genes in training (Section 4.3.5.4). The result of this experiment (Figure B.6) showed that an AUGUSTUS model trained on genes sampled from the reference annotation only slightly outperformed the model trained on the anchored genes. Conversely, training on highly conserved GeneMark-EP+ genes led to the lowest prediction accuracy.

Finally, we evaluated the effect of the number of anchored genes on the AUGUSTUS training (Section 4.3.5.4). As shown in Figure 4.7, low Sn and Sp values were observed for 500 genes. Both Sn and Sp significantly improved upon increasing the number of genes to

Table 4.4: *Ab initio* prediction accuracy of AUGUSTUS trained on (i) *All* genes predicted by GeneMark-EP+, and (ii) *Anchored* genes. The results for the first three species were generated with reference proteins from species outside a taxonomic family of a relevant species; for *D. rerio* we used proteins from species outside of the taxonomic order.
(*) When < 4000 anchored genes were available, additional genes were added in the descending order of their support by protein hints to reach 4000 genes (see Section 4.3.2). Particularly, this approach was used for *C. elegans*, which had 2,332 anchored genes.

	C. elegans		*A. thaliana*		*D. melanogaster*		*D. rerio*	
	All	Anchored*	All	Anchored	All	Anchored	All	Anchored
Gene Sn	38.6	**43.7**	52.8	**55.9**	51.7	**54.4**	15.0	**26.5**
Gene Sp	46.2	**50.9**	55.5	**56.9**	52.2	**55.7**	7.3	**13.1**
Gene F1	42.1	**47.0**	54.1	**56.4**	51.9	**55.0**	9.8	**17.5**
Exon Sn	**75.5**	75.1	**75.9**	75.7	68.5	**68.6**	68.3	**73.5**
Exon Sp	83.4	**86.2**	81.6	**83.2**	76.2	**80.5**	50.4	**63.0**
Exon F1	79.3	**80.3**	78.6	**79.3**	72.1	**74.1**	58.0	**67.8**

Table 4.5: The cumulative effect of new ProtHint hint types on the gene prediction accuracy of BRAKER2. The genome of *A. thaliana* and remote proteins (species of the same order excluded) were used on input

	ProtHint hints		
	high-confidence	+ non-high-confidence	+ CDSpart chains
Gene Sn	65.0	68.5	71.1
Gene Sp	63.6	66.0	67.0
Gene F1	64.3	67.2	69.0

1,000 and then kept increasing almost steadily when the number of genes increased from 1,000 to 8,000 genes. Based on this experiment, we chose 8,000 as the upper limit for the number of training genes and 4,000 as the minimum number of required training genes (as described in Section 4.3.2).

4.4.3 Impact of the novel ProtHint protein hints

Tables B.8 and 4.5 show the effect of using different types of ProtHint protein hints in BRAKER2, particularly the effect of integrating novel hint types that were not utilized by GeneMark-EP+ (see Section 4.3.3). For instance, in the genome of *A. thaliana* with remote proteins on input (species of the same order excluded), the use of non-high-confidence hints

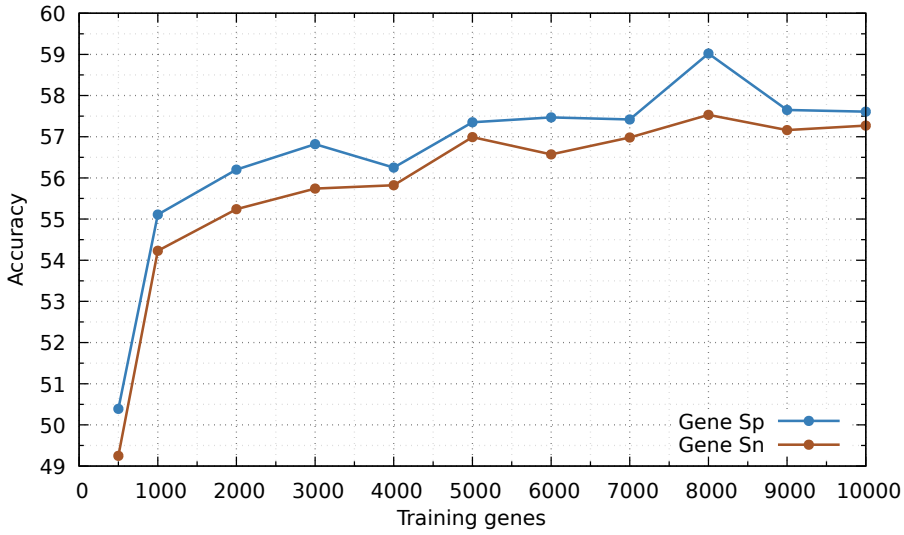

Figure 4.7: Dependence of AUGUSTUS *ab initio* gene prediction accuracy on the number of anchored genes in training. The experiment was done in the genome of *A. thaliana* and the supporting proteins outside of the Arabidopsis genus.

improved the gene-level F1 accuracy by ∼3 percentage points. The addition of chained CDSpart hints further improved the F1 accuracy by ∼2 percentage points (Table 4.5). Similar results for more species and protein databases are shown in the Appendix in Table B.8.

4.4.4 Prediction accuracy changes within the BRAKER2 pipeline

We observed a steady increase in the prediction accuracy upon moving from one to another step of the BRAKER2 pipeline (Table B.9). For instance, at the gene level, the F1 value for *D. melanogaster* supported by the largest protein database increased from GeneMark-ES to GeneMark-EP+ by 17.1 percentage points. Runs of AUGUSTUS with hints added 8.2 percentage points in the first iteration, and 1.1 percentage points in the second iteration. For the F1 values at the exon level, the numbers of increase were 8.8, 4.6, and 0.4 percentage points, respectively.

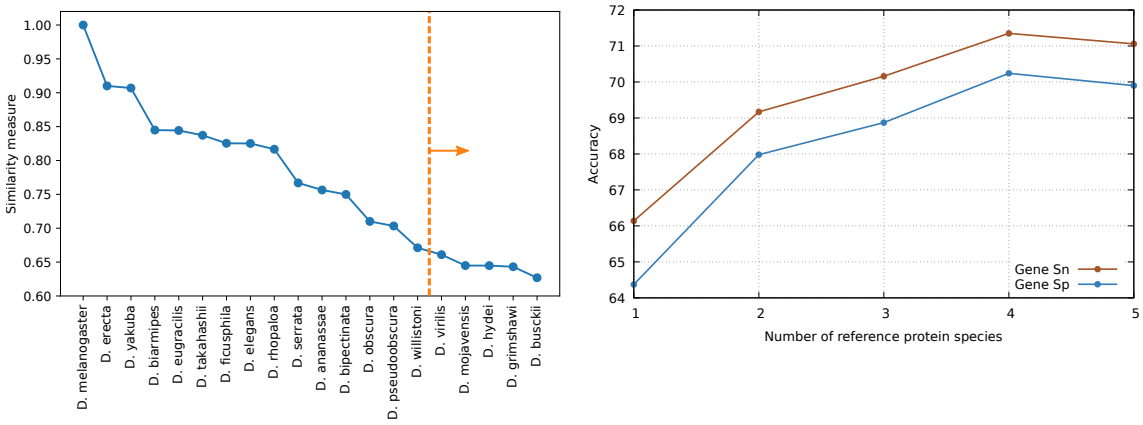

Figure 4.8: Change of BRAKER2 accuracy with the increasing number of species in the reference protein database. The left panel shows the evolutionary distance of species to *D. melanogaster* (see Section 4.3.5.5). The right panel shows the change in BRAKER2 accuracy with the increasing number of proteomes used on its input.

4.4.5 BRAKER2 prediction accuracy improves with the increasing number of species in the reference protein database

Figure 4.8 shows the results of the first experiment described in Section 4.3.5.5. The prediction accuracy of BRAKER2 was increasing steadily with the increasing number of the reference proteomes. We attribute the small decrease in the accuracy for five proteomes in training (instead of four) to stochastic effects in AUGUSTUS training.

The results of the second experiment are shown in Table 4.6. As expected, compared to using only Anopheles species, including all proteins of species outside of the taxonomic family led to a significant accuracy increase. Furthermore, even though the Anopheles species are in the same taxonomic order as *D. melanogaster*, using a larger number of proteins of species outside of *D. melanogaster*'s order also led to a slightly better gene prediction accuracy.

4.4.6 Comparison of BRAKER2 with MAKER2

The coordinates of genes predicted by MAKER2 in genomes of *A. thaliana*, *C. elegans* and *D. melanogaster* were compared to the annotations of the three genomes. When we used

Table 4.6: BRAKER2 prediction accuracy in *D. melanogaster* computed for several sets of input proteins: proteins of species (i) in the Anopheles genus, (ii) outside of *D. melanogaster*'s taxonomic family, and (iii) outside of *D. melanogaster*'s taxonomic order (Figure B.7).

	Proteins from		
	Anopheles genus	All outside of family	All outside of order
Gene Sn	60.7	66.3	61.1
Gene Sp	60.2	64.8	60.2
Gene F1	60.4	65.5	60.6

Table 4.7: Prediction accuracy of MAKER2 and BRAKER2.

	A. thaliana			*C. elegans*			*D. melanogaster*		
	MAKER2 with recommended protocol	MAKER2 with BRAKER2-like protocol	BRAKER2	MAKER2 with recommended protocol	MAKER2 with BRAKER2-like protocol	BRAKER2	MAKER2 with recommended protocol	MAKER2 with BRAKER2-like protocol	BRAKER2
Gene Sn	49.3	53.9	**70.6**	25.5	30.4	**43.7**	42.6	48.0	**60.0**
Gene Sp	42.1	55.6	**65.8**	22.1	38.9	**51.3**	31.1	50.3	**59.5**
Gene F1	45.4	54.7	**68.1**	23.7	34.1	**47.2**	35.9	49.2	**59.7**
Exon Sn	73.5	74.7	**80.6**	61.7	62.6	**71.9**	62.9	63.7	**71.3**
Exon Sp	72.6	83.0	**85.8**	64.5	81.4	**87.1**	58.7	76.0	**83.2**
Exon F1	73.0	78.6	**83.1**	63.1	70.8	**78.8**	60.7	69.3	**76.8**

the recommended MAKER2 protocol (Figure B.2A), the accuracy was significantly lower than the one of BRAKER2, which was run with the support of the same reference proteins. Particularly, gene-level F1 values were lower for *A. thaliana*, *C. elegans*, and *D. melanogaster* by 22.7, 23.5, and 23.8 percentage points, respectively (Table 4.7). Runs of MAKER2 with the second, "BRAKER2-like" protocol (Figure B.2B), resulted in reducing the gene-level F1 gaps between MAKER2 and BRAKER2 to 13.4, 13.1, and 10.5 percentage points, respectively (Table 4.7).

A detailed comparison of results obtained with the two MAKER2 protocols is shown in Table B.10. Training by BRAKER2-like protocol (training on genes predicted by GeneMark-ES and at least partially supported by protein alignments) produced better prediction accuracy than training directly from protein alignments as recommended by the default MAKER2 protocol (Table B.10). An improvement in the Sp values obtained as a result of using the BRAKER2-like protocol was largely related to the absence of SNAP, which generated an elevated number of false positive predictions. On the other hand, MAKER2

predictions with AUGUSTUS only were less accurate than with the combination of AUGUSTUS and GeneMark-ES (Table B.10).

With the exception of *C. elegans*, the predictions on unmasked sequences (Table B.11) showed an increase in prediction sensitivity and a decrease in specificity compared to the predictions on repeat-masked genomes. Still, the average gene-level Sn of MAKER2 on unmasked genomes (using the improved BRAKER2-like protocol) was 6.4 percentage points lower than that of BRAKER2, which was executed on repeat-masked sequences. For *C. elegans*, we observed an increase in both Sn and Sp when run on a masked genome. We attribute this behavior to MAKER2's hard-masking of all interspersed repeats (see Section 4.3.5.7), which in the case of *C. elegans* resulted in many predictions being corrupted due to repeat masking (11.9% of all annotated coding exons overlapped with sequences hard masked by MAKER2).

The runtimes of BRAKER2 and MAKER2 in our experiments were difficult to compare directly. We executed MAKER2 in the MPI mode [114] on a computational cluster with 96 CPUs. The runtime of MAKER2 ($\sim$10 h) using proteins from 10 species was comparable to the time needed for a run of BRAKER2 with proteins from 443 species executed on a single node with 8 CPUs.

4.5 Discussion

The goal of BRAKER2 was to achieve optimal integration of ProtHint, GeneMark-EP+, and AUGUSTUS—an integration which would combine the complementary strengths of these tools. This goal was successfully completed, making BRAKER2 a fully automated, state-of-the-art gene prediction pipeline with protein homology support; capable of utilizing proteins of any evolutionary distance to the genome of interest. In this section, we discuss the prediction accuracy of BRAKER2 and its dependence on the characteristics of the input protein database and the genome itself. Next, we highlight the main features contributing to BRAKER2's high prediction accuracy. Finally, we discuss how BRAKER2

compares with (i) RNA-Seq-based BRAKER1 and (ii), MAKER2, one of the most frequently used gene prediction pipelines.

4.5.1 BRAKER2 prediction accuracy analysis

4.5.1.1 The role of evolutionary distances and the total number of species in the protein reference set

The accuracy of a gene finding algorithm that utilizes cross-species proteins generally strongly depends on the evolutionary distance between the species of interest and reference proteins [57, 63, 64]. Indeed, for *A. thaliana*, *C. elegans*, and *D. melanogaster*, we saw that the accuracy steadily increased with the decreasing evolutionary distance of reference proteins (Figure 4.4). Nevertheless, even in tests with the most remotely related proteins, the accuracy of BRAKER2 was comparable to that of BRAKER1, which was supported by a large amount of RNA-Seq (Figure 4.4; a more detailed comparison with BRAKER1 is discussed later).

Another factor affecting the accuracy of BRAKER2 is the number of species whose proteins are used for generating the protein hints. For instance, the gene prediction accuracy observed for *A. thaliana* and *D. melanogaster*, which had more species involved in hints generation (Table 4.2), was higher than the accuracy for *C. elegans*, which had fewer species in each instance of the protein reference set (Figure 4.4). The benefits of increasing the number of species in the protein reference set were further demonstrated in a separate experiment (Figure 4.8). Moreover, we showed that the increase in the total number of species in the input protein set of BRAKER2 can compensate for the lack of close relatives. The use of many species outside of the *D. melanogaster*'s taxonomic order delivered better accuracy than the use of 14 Anopheles species within the order (Table 4.6).

It is important to also discuss situations in which a species with a sequenced and annotated genome exists at a very close distance to the genome of interest; i.e., when the average nucleotide identity computed for the two genomes is close to 100%. In such a case, a direct

gene annotation transfer could be a more efficient way of annotation compared to using BRAKER2; assuming that the reference annotation is of high quality. Otherwise, when the reference annotation is not trusted, the use of BRAKER2 would be a reasonable choice; BRAKER2's mechanism of hints generation accounts for the presence of errors in reference protein annotations.

4.5.1.2 The impact of the genome length and composition

We evaluated the accuracy of BRAKER2 on genomes that varied in length from 100 Mbp (*C. elegans*) to 1.4 Gbp (*X. tropicalis*). Notably, the exon level Sn computed on test sets of "reliable genes" (Section 4.3.5.2) remained at 80–90% for both shorter and longer genomes (Table 4.3). However, the gene level Sn, determined on the same test sets, showed a noticeable negative correlation with genome length. All genomes used in this study had relatively homogeneous nucleotide compositions. The current versions of the algorithms used in BRAKER2 employ a single set of species-specific models. The accuracy of BRAKER2 would drop down on genomes with heterogeneous composition, such as human (mammalian) or rice (grasses), where several models reflecting heterogeneous genome composition are necessary. This problem is addressed and solved in Chapter 5.

4.5.1.3 Completeness of BRAKER2 predictions

As a part of the accuracy assessment, we identified BRAKER2-predicted genes encoding proteins in species-specific BUSCO families. For a given genome, a fraction of BUSCO genes found provided a sensitivity estimate [118]. In all cases but *X. tropicalis*, the predicted sets of genes were comparable to or even more complete than the fractions of BUSCO genes identified in the available reference genome annotations (Figures B.3 to B.5). Notably, some of the BUSCO families were built from species within the same taxonomic order as the species of interest (e.g., Hemiptera order of *R. prolixus* or Solanales order for *S. lycopersicum*). At the same time, the only input to BRAKER2 were proteins of species

outside of the corresponding taxonomic order.

A lower level of BUSCO accuracy in *X. tropicalis* could be related to an insufficient number of external proteins. Removing the Anura taxonomic order from the OrthoDB partition left no proteins from the Amphibia taxonomic class among input proteins (Table 4.2). Also, a general cause for missing some of the BUSCO genes could be an inaccurate *de novo* repeat masking. For example, among the BUSCO genes missed by BRAKER2 in the *P. trichocarpa* genome, more than half were genes at least partially masked by long repeats (>1000 nt).

4.5.2 Sources of accuracy improvements

4.5.2.1 Automatic training

Previous attempts to automatically prepare (in the absence of transcripts) training sets for the training of supervised gene predictors were centered around the mapping of highly conserved cross-species proteins. However, these attempts were not successful due to the biases in the sets of the highly conserved genes [84]. We also indirectly observed this issue—when training AUGUSTUS on four different sets selected from GeneMark-EP+ predictions, the model trained on the most conserved genes achieved the lowest prediction accuracy (Figure B.6).

The new training approach of BRAKER2 uses cross-species protein conservation to predict protein hints and an *ab initio* gene prediction to connect the mapped hints into gene structures (i.e., GeneMark-EP+; Chapter 3). Compared to direct mapping of proteins or using highly conserved genes, this approach leads to a significant increase in the size of the training gene set supported by protein evidence. We argue that due to the sizes of the training sets, BRAKER2 does not suffer from the biases present in the sets of highly conserved genes (Figure 4.7). Furthermore, to achieve not just a large size, but also high quality of the training set, the selection of anchored training genes (Section 4.3.2) proved to be critical (Table 4.4).

Still, arguably, training on a sufficiently large set of manually curated gene structures (a supervised training) could outperform the automatic training implemented in BRAKER2. For instance, AUGUSTUS trained on a randomly selected set of annotated genes slightly outperformed AUGUSTUS trained by BRAKER2 (Figure B.6). Nonetheless, such an ideal condition (a random sampling from a 100% correct annotation) is an unlikely case when working with a novel genome.

4.5.2.2 Generation and integration of protein hints

BRAKER2 uses a novel approach for the preparation of protein hints. The protein mapping pipeline, ProtHint, predicts sets of hints with higher and lower confidence. All hints contribute to the generation of anchored genes used in training. GeneMark-EP+ enforces high-confidence hints in the prediction step. In turn, AUGUSTUS utilizes all hints at the prediction step along with information about the hints' connections within a putative transcript (CDSpart chains). While the use of high-confidence hints was already present in GeneMark-EP+ (Chapter 3), the use of non-high-confidence and chained hints is unique to BRAKER2.

The flexible use of hints leads to an increase in the accuracy of BRAKER2 (Tables B.8 and 4.5). Notably, since ProtHint is designed to generate accurate hints from proteins of remotely related species, BRAKER2 is a useful tool for the annotation of genomes of deeply branching species. Indeed, in all the results presented in this chapter, we observed that BRAKER2 performed well even when remote proteins (of species outside of the same order) were used on input (as already discussed in Section 4.5.1.1).

4.5.2.3 BRAKER2 iterations

The sensitivity of the ProtHint pipeline depends on the quality of seed genes predicted by GeneMark-ES. In particular, GeneMark-ES does not necessarily need to accurately predict the correct gene boundaries, it only needs to predict the correct gene loci; i.e., to have

high nucleotide-level sensitivity. Although GeneMark-ES has been shown to indeed have high nucleotide sensitivity [16], any *ab initio* gene finder may miss genes. In BRAKER2, these missed genes translate into missed protein hints to the corresponding genomic loci. The second iteration of BRAKER2 recovers hundreds of missed seed genes and leads to an increase in gene prediction accuracy (Table B.9). BRAKER2 could execute more than two iterations; however, we did not observe a significant increase in accuracy with three or more iterations.

4.5.3 Comparison of BRAKER2 with different gene finders

4.5.3.1 Comparison with BRAKER1

As we demonstrated, the accuracy of BRAKER2 depends on the number and evolutionary distance of reference proteins. In an analogous manner, the accuracy of BRAKER1 depends on the volume of the RNA-seq data. Experiments with BRAKER1 on genomes of *A. thaliana*, *C. elegans*, and *D. melanogaster* used RNA-Seq reads from NCBI SRA retrieved by VARUS; i.e. the non-redundant volumes of RNA-Seq reads from the maximum number of libraries available for each species.

When we used the largest number of supporting proteins for each species, only exempting proteins that originated from the tested genome, BRAKER2 was always more accurate than BRAKER1 with the above-described comprehensive RNA-Seq support. When using more remote proteins, the comparison between the accuracy of BRAKER2 and BRAKER1 was less clear cut (Figure 4.4); however, BRAKER2 with remote proteins was still more accurate than RNA-Seq-supported BRAKER1 in several of the tests.

Both BRAKER1 and BRAKER2 predicted a rather low number of annotated alternative isoforms (Table B.7). This is a result of a deliberate parameter setting in AUGUSTUS which aims to reduce the number of false positives. Particularly, AUGUSTUS ignored RNA-Seq or protein hints contradicting another hint with 10 times larger support. On the other hand, the reference genome annotations of the three species are relatively inclusive

in a sense of including potentially lowly expressed isoforms.

4.5.3.2 Comparison with MAKER2

The gap in accuracy between MAKER2 and BRAKER2 observed in our experiments was quite large despite our attempts to improve the default MAKER2 protocol (Table 4.7). This could be caused by differences in the methods of data preparation, training, processing repeats, ways of generating and selecting external evidence, connecting main elements of the pipelines, or combining the gene predictions into the final annotation. Therefore, we presented detailed descriptions of the protocols used for running MAKER2 (Section 4.3.5.7) as well as results obtained with different MAKER2 configurations (Tables B.10 and B.11).

MAKER2 uses the *ab initio* self-training algorithm GeneMark-ES, while BRAKER2 uses the more recent self-training GeneMark-EP+, which integrates protein hints into training and prediction. A more accurate training of AUGUSTUS is likely one of the important factors of BRAKER2's elevated accuracy. A comparison of protein hints generated by the two pipelines is difficult since BRAKER2 uses hints to splice sites and start/stop codons while MAKER2 uses hints to parts of exons.

The difference in the accuracy of BRAKER2 and MAKER2 is likely to be even larger in eukaryotic genomes with longer length; however, such a comparison is harder to make due to less accurate reference annotations. Since a comprehensive comparison of the two methods is not a goal of this thesis, the comparisons were limited to the three well-studied genomes.

Last but not least, the training of gene finders is not fully automated in MAKER2. Users have to execute the training steps manually, even though recommendations are given on the training protocols. On the other hand, BRAKER2 can be executed from start to finish with a single command, which is identical for any input genomic sequence.

4.5.3.3 Comparison with other gene finders

Several tools attempt accurate identification of gene structures in a novel genome by mapping homologous proteins (e.g., GenomeThreader [52], Scipio [122], or GeMoMa [57, 58]). This approach limits the gene discovery to genes of homologs present in the input protein set and the accuracy of these methods drops significantly with the increase of evolutionary distance between the species of interest and the reference proteins [57, 63, 64]. Another significant challenge, not addressed by the existing protein homology-based tools, is the processing of large volumes of proteins (which is understandable when focusing on a limited number of closely related species). Since the above-mentioned tools were not designed for the situation when the reference protein data do not contain proteins of closely related species, we did not compare their accuracy to BRAKER2.

4.6 Conclusion

BRAKER2 is a fully automated tool for gene prediction in novel eukaryotic genomes. BRAKER2 leverages information accumulated in protein databases, including proteins of large evolutionary distance to the species of interest. In tests on genomes of plants and animals, we observed that BRAKER2 delivered state-of-the-art annotation accuracy and was favorably compared to already existing tools.

4.7 Availability

BRAKER2 is available at https://github.com/Gaius-Augustus/BRAKER. All additional scripts and data used to generate figures and tables in this chapter are available at https://github.com/gatech-genemark/BRAKER2-exp.

BRAKER2 does not require significant computational resources; for instance, in the case of *D. melanogaster*, $\sim$2.6 million proteins were processed in $\sim$3 h on a single node with 8 CPU cores; the overall BRAKER2 runtime being $\sim$10 hours.

CHAPTER 5

GENEMARK-ETP+: AUTOMATIC INTEGRATION OF GENOMIC, TRANSCRIPTOMIC, AND PROTEIN DATA FOR GENE PREDICTION IN EUKARYOTIC GENOMES

Abstract

An integrative method of gene prediction in eukaryotic genomes has to solve the complex task of parsing the genome into coding and non-coding regions in agreement with extrinsic evidence at transcript and protein levels. We present GeneMark-ETP+, a new addition to the family of the GeneMark eukaryotic gene finders with unsupervised parameter training. GeneMark-ETP+ utilizes transcriptomic, protein homology, and intrinsic data evidence sources throughout all stages of the algorithm's training and gene prediction. Both the transcript and protein evidence have an uneven distribution across a genome. Therefore, GeneMark-ETP+ proceeds, first, with the identification of genes in loci where extrinsic data density is sufficient for gene identification with high confidence, and, second, with gene finding in fragments between the high-confidence genes. The performance of GeneMark-ETP+ was favorably compared with methods using a single type of extrinsic evidence such as GeneMark-ET, GeneMark-EP+, BRAKER1, and BRAKER2. Furthermore, GeneMark-ETP+ achieved higher prediction accuracy than TSEBRA, a recently developed algorithm combining the predictions of RNA-Seq-based BRAKER1 and protein homology-based BRAKER2.

5.1 Introduction

Gene prediction algorithms generally utilize data from one or more of the following sources (Section 2.2): (i) protein homology, (ii) transcriptomic evidence, and (iii) intrinsic statis-

tical patterns of the genomic sequence itself. Strictly homology-based approaches (e.g., exonerate [51], GenomeThreader [52], or ProSplign [53]) and exclusively transcriptomic-based ones (e.g., StringTie [33, 34], PsiCLASS [35], or Cufflinks [36]) are limited to the discovery of similar enough and sufficiently expressed genes, respectively. To correctly predict novel/remotely homologous and weakly expressed genes, prediction algorithms need to utilize intrinsic sequence features such as splice site motifs, intron/exon length distributions, codon usage, etc. Several early-developed algorithms (e.g., Genie [10], GENSCAN [14], GeneID [123], SNAP [13], AUGUSTUS [12]), so-called *ab initio* methods, rely solely on these intrinsic features. Unfortunately, the accuracy of such methods is far from perfect [19–22], especially for large eukaryotic genomes, because parts of the gene structures often lack strong motifs that can be accurately predicted [18]. Therefore, modern gene finders rely on intrinsic as well as proteomic and/or transcriptomic information sources. GeneMark-EP+ (Chapter 3) and BRAKER2 (Chapter 4) are the most recent examples of algorithms integrating protein homology and *ab initio* components. On the other hand, GeneMark-ET [48] and BRAKER1 [47] are examples of algorithms integrating RNA-Seq and intrinsic data sources. Notably, GeMoMa [58] utilizes RNA-Seq data to enhance homology-based predictions. However, GeMoMa lacks an *ab initio* component and is limited to the annotation of closely related genomes only.

The simultaneous integration of all three information sources (protein homology, transcriptomic, and genomic) remains an open problem. The majority of tools integrating all the information (e.g., FINDER [71], LoReAn [72], GAAP [73], IPred [74], Evigan [75], EVidenceModeler [76], JIGSAW [77], Combiner [78], or GAZE [79]) work as combiners: Their approach is to first generate multiple independent *ab initio*-, protein homology-, and transcriptomic-based predictions and subsequently combine them into a prediction that is, on average, more accurate than any input source. Thus, the distinct evidence streams are only integrated at a "post-processing" step of the gene prediction process. A better prediction accuracy could potentially be achieved by integrating the different information sources

in all prediction steps as well as during the training of prediction models.

In this chapter, we introduce GeneMark-ETP+, a tool combining transcriptomic, protein homology, and intrinsic data sources throughout *all* stages of the algorithm's training and gene prediction. GeneMark-ETP+ facilitates this integration by, among other things, creating a novel method for simultaneous gene prediction in transcripts (assembled from RNA-Seq) and genomic DNA. Importantly, the training of GeneMark-ETP+ is fully unsupervised and the protein homology evidence integration utilizes proteins of any evolutionary distance, including remote homologs (using techniques developed in Chapters 3 and 4). Last but not least, GeneMark-ETP+ also focuses on a comprehensive integration of repeat annotations, an important information source that is often not given enough attention in current gene prediction algorithms ([3, 124–126]). All over, due to its ability to efficiently integrate all major gene prediction information streams in an unsupervised manner, GeneMark-ETP+ offers an important step toward fully automated and accurate eukaryotic gene prediction.

We assessed the prediction accuracy of GeneMark-ETP+ on seven genomes representing compact GC-homogeneous as well as large GC-heterogeneous eukaryotic genomes. We compared the prediction accuracy of GeneMark-ETP+ with GeneMark-ET, GeneMark-EP+, and the best theoretical combination of predictions made by these two tools. Further, we compared GeneMark-ETP+ with TSEBRA [70], a recently published combiner that combines the results of RNA-Seq-based BRAKER1 and homology-based BRAKER2.

5.2 Materials

For the assessment of GeneMark-ETP+, we selected seven genomes representing diverse eukaryotic clades (Tables C.1 and 5.1). Based on their genomic organizations, the selected genomes can be split into three groups. The first group, genomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*, represented early-sequenced model organisms with relatively short GC-homogeneous genomes. The remain-

Table 5.1: Genomes used for the assessment of GeneMark-ETP+ accuracy. The numbers in parentheses characterize the reliable annotation subsets. Introns per gene were computed as a weighted average: the # of introns in each gene G was inversely weighted by the # of alternative transcripts in G. Without this adjustment, the average would be skewed towards genes with many annotated isoforms.

Species	Genome length (Mb)	Reference annotation statistics		
		# coding genes	# coding transcripts	introns per gene
C. elegans (roundworm)	100	19,969	28,544	4.8
A. thaliana (thale cress)	119	27,445	40,828	4.0
D. melanogaster (fruit fly)	138	13,951	22,395	2.8
S. lycopersicum (tomato)	807	25,158 (15,138)	31,911 (15,150)	4.4 (4.3)
D. rerio (zebrafish)	1,345	25,611 (17,894)	42,934 (19,978)	8.4 (8.4)
G. gallus (chicken)	1,050	17,279 (10,736)	38,534 (12,733)	9.0 (9.2)
M. musculus (mouse)	2,723	22,405 (16,531)	58,318 (20,708)	6.0 (8.6)

ing two groups, both containing larger genomes, represented GC-homogenous (*Solanum lycopersicum*, *Danio rerio*) and GC-heterogeneous (*Gallus gallus*, *Mus musculus*) genomes. In all genomic datasets, contigs not assigned to any chromosome and the genomes of organelles were excluded from the analysis.

We used OrthoDB v10.1 protein database [66, 67] to prepare the input sets of cross-species proteins for each species in the test set (Table C.2). Each protein set was generated in the same way as described earlier in Sections 3.2.1 and 4.2.1. First, all proteins from a large clade corresponding to the query species were used to create an initial protein set S. Next, two smaller sets were created by removing from S (i) all proteins of the query species itself, and (ii) all proteins from species of the same taxonomic order. The reduced sets of type (ii) were generated to simulate large evolutionary distances between the query species and input proteins, which could be expected in practical situations of running GeneMark-ETP+ on a newly sequenced genome.

RNA-seq libraries of the Illumina paired reads were selected from the NCBI SRA database [93]. The length of reads varied from 75 to 151 nt and the total size of all the reads varied from 9 Gb in *D. melanogaster* to 83 Gb in *M. musculus* (Table C.3).

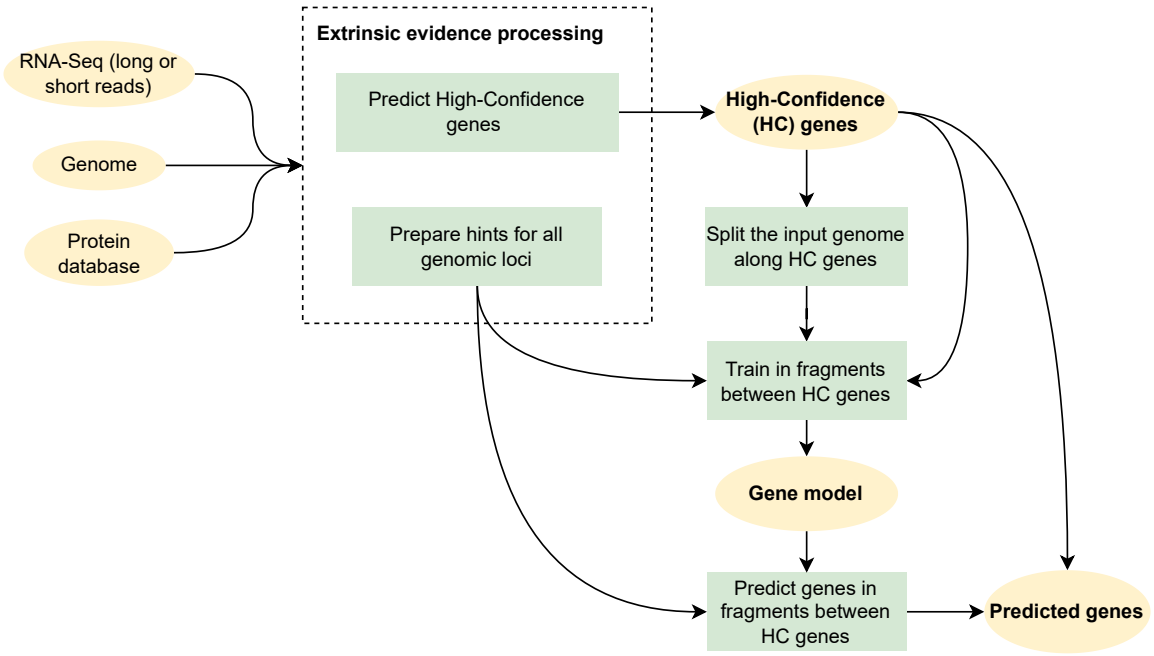

Figure 5.1: A high-level overview of GeneMark-ETP+.

5.3 Methods

5.3.1 Overview of GeneMark-ETP+

GeneMark-ETP+ consists of three major parts. First, GeneMark-ETP+ infers a set of reliable exon-intron structures, so-called high-confidence (HC) genes, directly from the transcriptomic and protein evidence. Second, the high-confidence genes are used to train a species-specific model and to split the genomic sequence into non-overlapping fragments. Finally, the trained model is used to predict genes in the fragments between high-confidence genes. Available sources of transcript and protein evidence, even less reliable ones, such as the alignments of remotely homologous proteins, are used in all three steps. A high-level GeneMark-ETP+ diagram is shown in Figure 5.1.

Notably, GeneMark-ETP+ predicts genes in the assembled transcriptome as well as in the genomic DNA and combines these two prediction sets into the final genome annotation. Two distinct types of generalized hidden Markov models (GHMM) are used in the process.

A GHMM with an intron model (described in GeneMark-ES [16]) is used to predict genes in the genome and a GHMM with no intron model (described in [43]) is used to predict genes in transcripts. The training of the latter model (with no intron model) is done by an unsupervised GeneMarkS-T [43]. The unsupervised training of the genomic GHMM model (with an intron model) is done by a new procedure that makes use of the transcriptome predictions.

In the previously described GeneMark-EP+ (Chapter 3), as well as in GeneMark-ES/ET [16, 17, 48], the genomic GHMM was trained in an iterative unsupervised (or semi-supervised) manner. In GeneMark-ETP+, depending on the volume of external data, the iterative training procedure is not always necessary; in some situations, the GHMM can be trained directly from the predictions made in transcripts.

5.3.2 Prediction of High-Confidence genes

The overview of the high-confidence gene prediction process is shown in Figure 5.2. Briefly, transcripts are first assembled from RNA-Seq reads. Next, genes are predicted in the transcripts with GeneMarkS-T (GMS-T) [43]. The raw GeneMarkS-T output may contain incorrect predictions stemming from errors in the transcript assembly and errors made by GMS-T itself. Furthermore, a large portion of correct predictions may be incomplete due to the presence of incomplete transcripts. The shares of the incomplete and incorrect predictions depend on the quality of the RNA input and other factors; thus they significantly vary between inputs. GeneMark-ETP+ employs a series of classification and filtering steps to categorize complete and incomplete predictions and to remove false predictions from the GMS-T output; thus creating a highly reliable set of high-confidence genes.

5.3.2.1 Transcript assembly and gene prediction in transcripts with GeneMarkS-T

To create transcript assemblies, RNA-Seq reads are mapped to the genome with HISAT2 [32] and subsequently assembled into transcripts with StringTie [33]. In this process, each

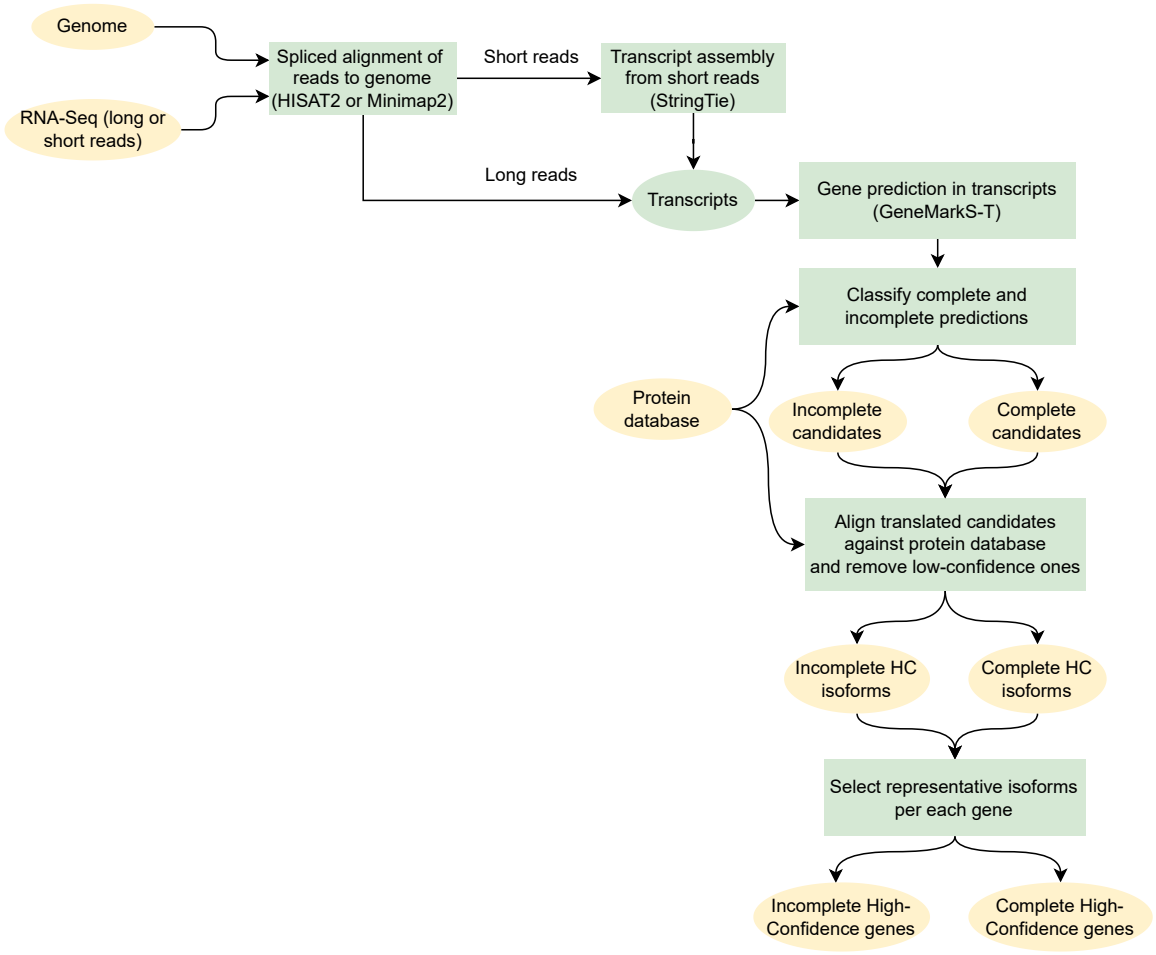

Figure 5.2: A high-level overview of the high-confidence gene prediction.

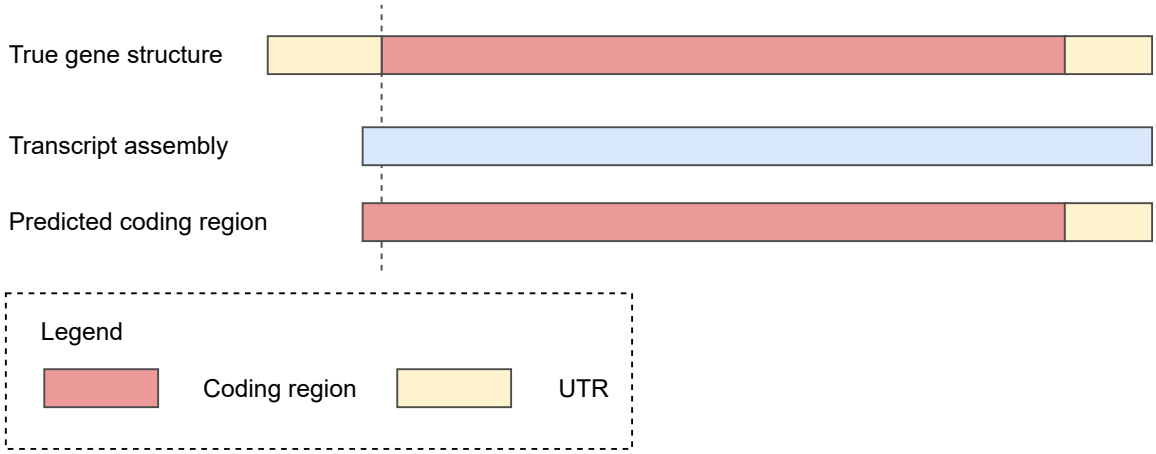

Figure 5.3: Example of an incorrect incomplete coding gene prediction. Although the 5' UTR of the assembled transcript is incomplete, the assembly contains the full coding region. However, due to the short length of the available non-coding sequence, the predicted coding region was incorrectly extended to the 5' end of the sequence.

RNA-Seq library is first aligned and assembled separately. At the end, the individual assemblies are merged into a single transcriptome assembly with StringTie. Once transcripts are assembled, GeneMarkS-T, an algorithm with unsupervised training, is used to predict coding regions in the transcripts.

If long RNA reads are available, they can be directly mapped to the genome with Minimap2 [40], without the need for the assembly step. However, the experiments in this thesis only describe situations with short RNA-Seq reads on input.

5.3.2.2 Classification of predictions as complete and incomplete

GMS-T often predicts that a truly complete transcript with a short 5' UTR is incomplete—by missing the UTR and extending the coding region to the 5' end of the sequence (as illustrated in Figure 5.3). This makes the incomplete predictions highly unreliable. We reduce the error rate by the following procedure.

Each 5' incomplete coding region predicted by GMS-T is first shortened to the first in-frame ATG start codon. If there is no such start codon, the prediction is classified as incomplete. Both initially predicted incomplete coding region and the shortened one are

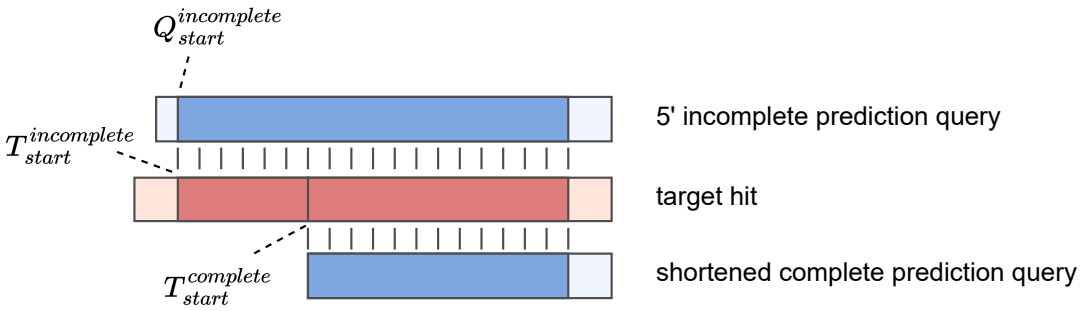

Figure 5.4: The alignment features used to classify complete and incomplete GMS-T predictions.

translated into proteins and searched for similarities against the input protein database by DIAMOND [87] in the BLASTp mode. Next, up to 25 best alignments are selected from both searches. If at least one pair of these alignments, made against the same target protein from the database, shows better support for the longer incomplete sequence, the GMS-T prediction is classified as incomplete (i.e., the classification agrees with the original GMS-T prediction). Otherwise, the prediction is classified as complete, and it is subsequently represented by its shortened version.

To compute the support score, we use the following features (illustrated in Figure 5.4):

- $Q_{start}^{incomplete}$ – the position of the start of the alignment in the incomplete query protein.
- $T_{start}^{incomplete}, T_{start}^{complete}$ – the positions of the start of the alignment in the target protein when aligned against the incomplete and complete candidate, respectively.
- $AAI^{incomplete}, AAI^{complete}$ – the percentages of amino acid identities in the alignments of the incomplete and complete candidates, respectively.

If the inequality in Equation (5.1) is satisfied, the alignment supports the incomplete candidate.

$$\left(T_{start}^{complete} - T_{start}^{incomplete}\right) - \left(Q_{start}^{incomplete} - 1\right) + \ln\left(\frac{AAI^{incomplete}}{AAI^{complete}}\right)^{1000} > 0 \qquad (5.1)$$

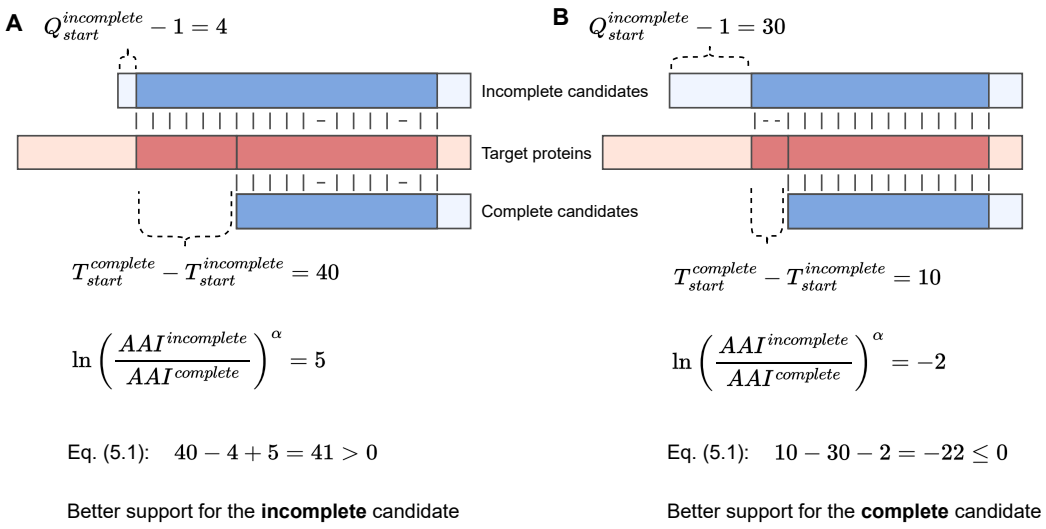

Figure 5.5: Examples of GMS-T predictions classified as incomplete and complete, respectively.

Examples of both complete and incomplete classifications, which illustrate how each of the features contributes to the support score, are shown in Figure 5.5. Further details describing how these features were derived are given in the Discussion.

Regarding 3' incomplete predictions (unambiguously defined by the lack of a stop codon)—these are excluded from the high-confidence output. Finally, genes predicted as complete by GMS-T are not further re-classified—they are always treated as complete in GeneMark-ETP+. See the Discussion for the rationale behind these design decisions.

5.3.2.3 Selection of high-confidence GMS-T gene predictions

After the above-described classification, both complete and incomplete GMS-T candidates are translated to proteins and aligned against the protein database with DIAMOND in the BLASTp mode. Then, the following rules are applied to select high-confidence gene predictions.

Complete high-confidence predictions supported by proteins

For each complete GMS-T prediction candidate, the following features are extracted from

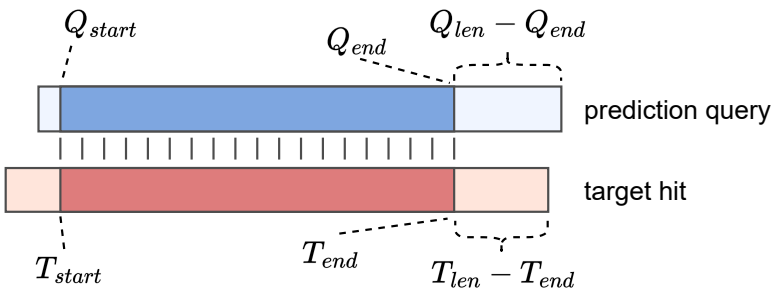

Figure 5.6: The alignment features used to classify complete high-confidence genes supported by proteins.

up to 25 top scoring target protein alignments against the complete query (illustrated in Figure 5.6):

- Q_{start}, Q_{end} – the positions of the start and end of the alignment in the query protein, respectively.
- T_{start}, T_{end} – the positions of the start and end of the alignment in the target protein, respectively.
- Q_{len}, T_{len} – the length of the query protein and the aligned target, respectively.

If the alignment of any of the target proteins satisfies the condition in Equation (5.2), the GMS-T candidate is classified as a high-confidence one.

$$(Q_{start} - T_{start} \leq 5) \wedge ((Q_{len} - Q_{end}) - (T_{len} - T_{end}) \leq 20) \tag{5.2}$$

As before, more details explaining the design of Equation (5.2) are given in the Discussion.

Incomplete high-confidence predictions supported by proteins

For an incomplete gene prediction to be classified as high-confidence, its C-terminus must be supported by at least one protein alignment (as in Equation (5.2)). Since the support for the incomplete N-terminus is implied by Equation (5.1), all incomplete candidates with

the C-terminus support are classified as high-confidence. However, if the best-scoring protein alignment does not cover the incomplete prediction from its start ($Q_{start} \neq 1$), the incomplete high-confidence gene prediction is shortened to the first in-frame ATG start.

High-confidence predictions supported by intrinsic evidence

The GMS-T gene predictions that have no significant hits on protein level or do not satisfy Equation (5.2) are filtered as follows. First, all isoforms to HC genes already selected in the previous steps are removed from consideration. Next, only one representative (having the longest protein-coding region) is selected per a gene with multiple unsupported isoforms. To be classified as a high-confidence gene, the selected representative must: (i) have length ≥ 300 nt, (ii) have GMS-T log-odds score > 50, (iii) be classified as complete, (iv) have in-frame stop codon in the 5' UTR, and (v) create no conflict with some other viable gene predicted in the same locus. The last condition is verified by mapping the GMS-T-predicted gene to genomic DNA and comparing the resulting exon-intron structure with a potentially conflicting ProtHint prediction. The exact algorithm for checking this last condition is described in the Appendix in Section C.1.

5.3.2.4 Adjustment of GMS-T predictions creating less than longest ORF

Complete GMS-T gene predictions that could be extended by changing the position of the start codon are subjected to additional analysis. For each such gene, GeneMark-ETP+ generates an alternative by extension to the longest open reading frame (ORF). These alternatives are subjected to the filtering procedure described above. If the alternative satisfies Equation (5.2), it is used instead of the original prediction.

5.3.2.5 Alternative high-confidence isoforms

Selected high-confidence gene predictions may include alternative isoforms of the same gene. Note that the alternative isoforms are considered only for gene predictions supported

by extrinsic evidence (vs intrinsic evidence, see Section 5.3.2.3). GeneMark-ETP+ selects a subset of reliable protein-supported isoforms in the following way.

Let $I^g_{complete}$ be a set of all complete isoforms of gene g and $I^g_{incomplete}$ a set of all its incomplete isoforms. Each isoform i is assigned a score $s(i)$—the *bitscore* of its best protein hit in the protein database.

To select complete alternative isoforms, first, the maximum of scores of all complete isoforms is computed for each gene g:

$$s(g_{complete}) = \max_{i \in I^g_{complete}} s(i) \tag{5.3}$$

The score of each isoform must satisfy the condition in Equation (5.4); otherwise, the isoform is removed.

$$s(i) \geq 0.8 \times s(g_{complete}) \quad (i \in I^g_{complete}) \tag{5.4}$$

As a result, several isoforms can represent a complete high-confidence gene.

Conversely, only one representative, best supported by a protein alignment, is selected among incomplete alternative predictions. Again, the maximum among scores of all incomplete isoforms is computed for each gene g:

$$s(g_{incomplete}) = \max_{i \in I^g_{incomplete}} s(i) \tag{5.5}$$

The transcript with the score corresponding to this maximum (Equation (5.6)) is selected to represent the incomplete high-confidence gene.

$$s(i) = s(g_{incomplete}) \quad (i \in I^g_{incomplete}) \tag{5.6}$$

After this processing, a gene g may have both complete and incomplete HC isoforms. The

incomplete prediction is removed from the set if condition in Equation (5.7) is fulfilled. Otherwise, the HC gene is represented by the single incomplete prediction, and all the complete isoforms are removed.

$$s(g_{complete}) \geq s(g_{incomplete}) \tag{5.7}$$

5.3.3 Genomic model training and genome segmentation

5.3.3.1 Training of a genomic GHMM

The training of a genomic generalized hidden Markov model (GHMM), which is used to predict genes in the genomic sequence, is done based on the predicted high-confidence genes as follows. For training, the set of complete HC genes predicted in the assembled transcripts is reduced by selecting a single isoform of each HC gene (the one with the longest protein-coding region). These isoforms are mapped to the genomic DNA. Next, genomic loci and gene structures corresponding to these mapped genes are used to train the GHMM.

Before the actual training starts, the average GC content of each training gene is calculated. If the genes' GC content distribution is such that more than 75% of genes fall within a 10%-wide GC bin, the species' genome is considered to be GC-homogeneous. Otherwise, it is considered to be GC-heterogeneous.

For genomes classified as GC-homogeneous, all selected HC training genes are used for the estimation of model parameters. If the training set is large enough ($>$ 4,000 genes), the training is concluded (Figure 5.7). Otherwise, an iterative training procedure, akin to the one used in GeneMark-EP+ (Section 3.3.3.1), is executed (see Section 5.3.3.3).

For genomes classified as GC-heterogeneous, GeneMark-ETP+ splits the training set of HC genes into three GC-bins: low, medium, and high. The width of the medium GC bin is set to 10%. The bin's location is selected to include the largest possible number of genes from the HC training gene set. The low and medium GC bins are then unambiguously

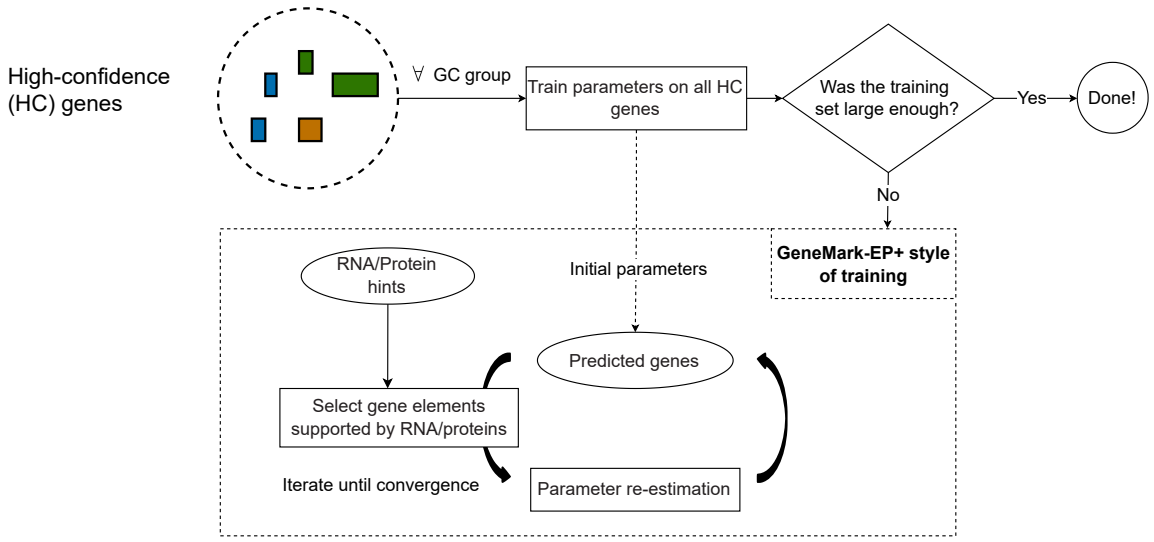

Figure 5.7: Training of the genomic generalized hidden Markov model (GHMM).

defined with respect to the medium GC bin. The training of the GHMM is performed inside each GC bin using the same method as described above for the GC-homogeneous genomes.

5.3.3.2 Genome segmentation

GeneMark-ETP+'s step of genome segmentation, which is used in the extended training as well as in the final prediction, is an important feature not present in the previous GeneMark versions. A genome is segmented by the locations of all mapped high-confidence genes, including the incomplete ones (Figure 5.8). The loci corresponding to the complete high-confidence genes are used in the training of the GHMM (Section 5.3.3.1). In the remaining segments, so-called *non-HC-segments*, the trained GHMM is used to predict genes (further described in Section 5.3.4).

In the GC-heterogeneous genomes, GeneMark-ETP+ calculates the GC content of each non-HC-segment and assigns segments to one of the GC bins according to the GC bin classification from the model training step (Figure 5.8). Gene prediction and the extended training are then executed separately in each bin using GC-bin-specific models.

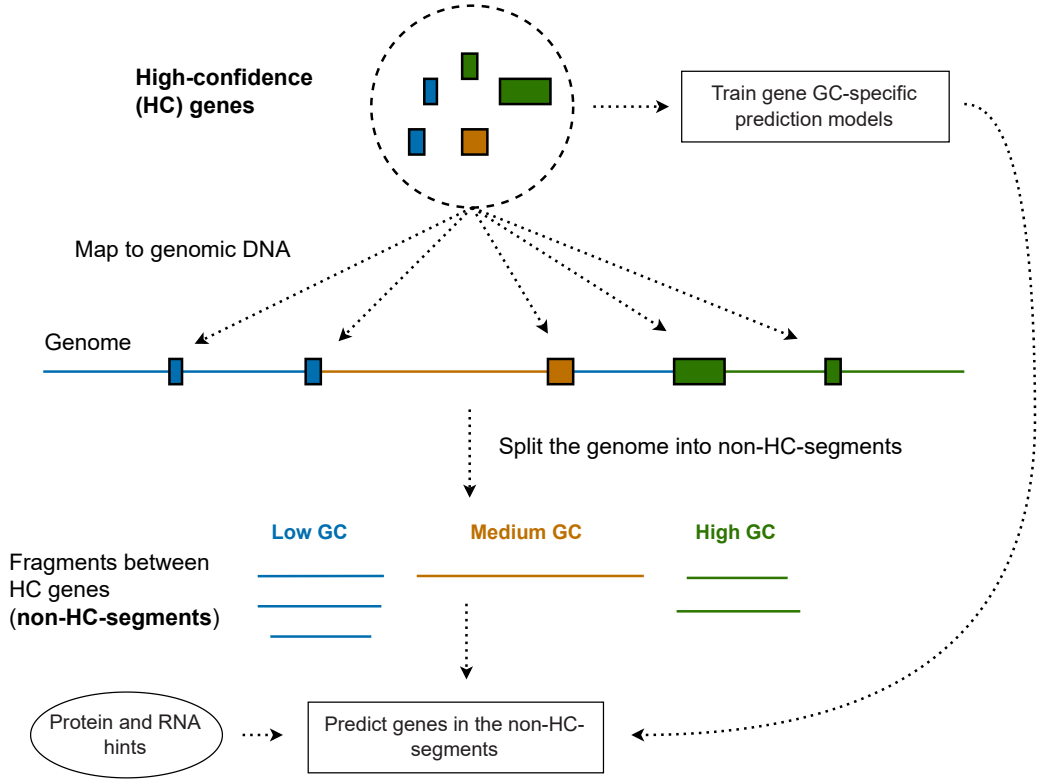

Figure 5.8: Splitting of the genome into non-HC-segments between high-confidence genes.

5.3.3.3 *Extended training*

The extended training, executed when the set of training high-confidence genes contains $\leq$ 4,000 genes, is implemented in the same way as the iterative anchored training described in Section 3.3.3.1. First, an initial gene prediction in non-HC-segments is created using a GHMM trained from high-confidence genes. Then, training hints (mapped from RNA-Seq and created by ProtHint), are used to iteratively extend the training set of high-confidence genes by adding anchored elements from predictions in non-HC-segments (Figure 5.7).

5.3.3.4 *Repeat penalty and its estimation*

GeneMark-ETP+ changes the probability of a putative protein-coding sequence overlapping a repeat by using a penalty q (n is the length of the overlap):

$$P(seq|coding\ state\ overlapping\ repeat) = \frac{P(seq|coding\ state)}{q^n} \tag{5.8}$$

The optimal (i.e., leading to the highest gene prediction accuracy) repeat penalty value may depend on the species and the repeat annotation; therefore, GeneMark-ETP+ estimates the optimal q during its unsupervised training. Particularly, the estimation algorithm attempts to find the highest penalty value that does not disrupt correct predictions. The estimation routine uses the already trained generalized hidden Markov model (GHMM), the predicted high-confidence genes, genomic DNA, and predicted repeat coordinates. The input for this procedure is a set of genomic loci containing mapped high-confidence genes, extended by 1000 nt on both ends. The GHMM is run with different q values to predict genes in this sequence set (with no extrinsic evidence other than repeat annotation). The accuracy of the predictions is evaluated on the mapped HC genes. In the first stage, the penalty estimation algorithm finds q that maximizes e_{max}, the number of correctly predicted exons. This often corresponds to a value q close to 1. In the second stage, the algorithm looks for the highest

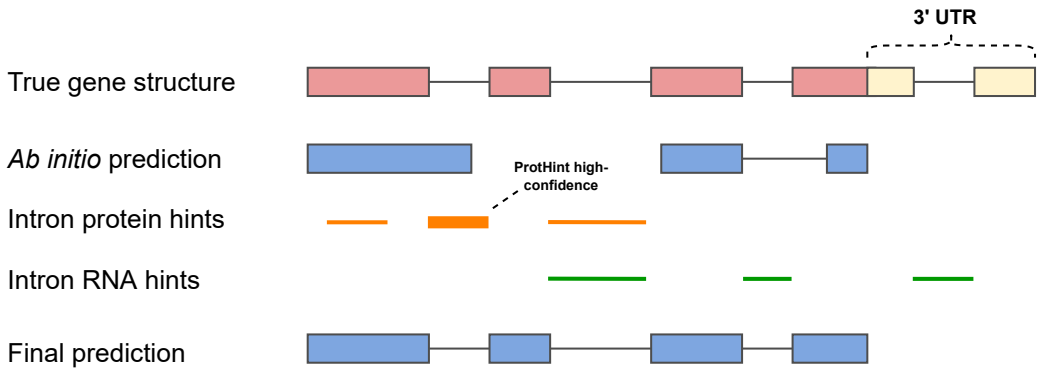

Figure 5.9: Integration of extrinsic evidence into the GeneMark-ETP+ gene predictions in non-HC-segments between mapped HC genes.

q with which the model correctly predicts $\geq 0.998 \times e_{max}$ exons. This q is the predicted optimal value. To make this procedure fast, the penalty search space is explored using an approach similar to simulated annealing [127].

5.3.4 Gene predictions in non-HC-segments of genomic DNA

In the last stage of GeneMark-ETP+, the generalized hidden Markov models trained in the previous step are used to create preliminary predictions in the non-HC-segments (Section 5.3.3.2). These predictions serve as seed genes for the ProtHint protein mapping pipeline (Section 3.3.2). All GMS-T gene predictions excluded from the HC set are also used as ProtHint's seed genes. After the run of ProtHint, protein, as well as RNA-Seq evidence (generated during RNA-Seq mapping with HISAT2), are ready to be used for the final predictions in non-HC-segments between mapped HC genes.

5.3.4.1 Integration of genomic, transcriptomic, and protein evidence

Both evidence types (proteins and transcripts) are integrated into the final *ab initio* prediction (Figure 5.9) to boost the accuracy of the generalized hidden Markov models. This integration is done as follows.

1. Hints predicted independently from both proteins and RNA-Seq, are enforced into

the predicted gene structures (as in Section 3.3.3.2).

2. Hints in loci covered exclusively by protein evidence are enforced if scored high by ProtHint.

3. *Ab initio* predictions in loci covered exclusively by RNA-Seq evidence are enforced to agree with RNA-Seq mapping in all cases where it can be confidently concluded that the evidence does not originate from UTRs or non-coding genes. Specifically, only RNA-Seq introns that conflict with a pure *ab initio* intron (i.e., intron not supported by any protein or RNA-Seq hints) are enforced.

4. Finally, all incomplete HC genes, identified in the previous step (Section 5.3.2.3), are extended if possible to full predictions using the GHMM.

The final predictions in the non-HC-segments, together with the complete HC genes themselves (mapped to genomic DNA), constitute the final set of genes predicted by GeneMark-ETP+.

5.3.4.2 Filtering of pure ab initio predictions

The final gene predictions in non-HC-segments can be split into two non-overlapping sets: evidence-supported and pure *ab initio* predictions. The evidence supported genes must have at least one element of their gene structure (either intron, start, or a stop) supported by the protein or transcriptomic evidence. We observed that in larger genomes, the fraction of correct purely *ab initio* predictions steadily decreased with the genome size. Therefore, GeneMark-ETP+ offers two types of outputs for genomes larger than 300 Mbp—the first with the full set of predictions and the second with pure *ab initio* predictions removed. We used the second, reduced, output in the accuracy tests done for the four larger genomes.

5.3.5 Methods related to algorithm assessment

This section describes the design of methods that were used to evaluate aspects of GeneMark-ETP+; other than the standard accuracy assessment described in Section 2.4. In all evaluations, regions of annotated pseudogenes were excluded from comparisons.

5.3.5.1 Repeat masking

To identify repetitive sequences, we used RepeatModeler2 [128] and RepeatMasker [129]. First, a repeat library was generated *de novo* using RepeatModeler2. Repeat sequences—interspersed and tandem repeats—were subsequently found and soft-masked using RepeatMasker.

To assess the efficiency of the algorithm for automatic estimation of repeat masking penalties (Section 5.3.3.4), we conducted the following experiment. We executed the final prediction step of GeneMark-ETP+ with varying repeat penalties and calculated the prediction accuracy with each penalty setting. These predictions were executed (and evaluated) in the non-HC-segments (Section 5.3.3.2) between HC genes as HC genes themselves are not affected by the repeat penalty. In order to illustrate how the masking estimation procedure operates, we also plotted how the fraction of correctly predicted HC exons during penalty estimation (see Section 5.3.3.4) depended on the varying penalty values.

5.3.5.2 Selection of reliable annotation subsets

Since curated genome annotations of *A. thaliana*, *C. elegans*, and *D. melanogaster* have been updated multiple times, we considered their current complete annotations as "gold standards". Thus, all accuracy comparisons involving these genomes were done with respect to the full complement of annotated genes. The reference annotations of the other four genomes are arguably less trustworthy [3]. Therefore, we prepared reliable annotation subsets containing coding transcripts that have identical annotations in two different sources; see Table C.1 for the annotation sources used for each genome. Overall statistics

describing the reliable genes are shown in parentheses in Table 5.1. The prediction sensitivity estimates for the four large genomes were computed against these reliable annotation subsets.

5.3.5.3 Optimal combination of GeneMark-ET and GeneMark-EP+

We compared the accuracy of GeneMark-ETP+ with the accuracy corresponding to the best possible (hypothetical) combination of gene sets predicted by RNA-Seq-based GeneMark-ET and protein-based GeneMark-EP+. To make such a comparison, we prepared two combinations of GeneMark-ET and -EP+ predictions: their union (U) and intersection (I). The intersection only contained genes with identical gene structures. Arguably, set U represents the most comprehensive combination while I represents the most reliable one. Thus, the sensitivity and specificity of the optimal combination were set to the Sn of U and the Sp of I.

5.3.5.4 Running BRAKER1, BRAKER2, and TSEBRA

We ran RNA-Seq-based BRAKER1 [47] and protein-homology-based BRAKER2 (Chapter 4) with the same sets of input RNA-Seq libraries and protein databases, respectively, as the ones used by GeneMark-ETP+. See Chapter 4 for more details about BRAKER1 and (especially) BRAKER2. Next, we combined the results of BRAKER1 and BRAKER2 with TSEBRA [70]—a tool that finds an optimal combination of predictions made by BRAKER1 and BRAKER2 and generates a gene prediction set supported by both RNA-Seq and homologous protein evidence. In a recently published paper [70], TSEBRA was shown to achieve higher accuracy than (i) either BRAKER1 or BRAKER2 running alone, and (ii) EVidenceModeler [76], one of the most prominent combiner tools.

5.4 Results

We evaluated the accuracy of GeneMark-ETP+ for seven genomes representing diverse taxonomic clades and genomic organizations (Section 5.2), using the accuracy evaluation methods presented in Section 5.3.5. The experiments for each species were conducted with two different protein databases (Section 5.2): (i) a database where only the proteins of the query species itself were removed, and (ii) a database where all proteins from species of the same taxonomic order were removed. To demonstrate the accuracy in the absence of closely related cross-species proteins, most results shown in this section were generated with the latter protein database on input (all results are shown in the Appendix).

5.4.1 Accuracy assessment of GeneMark-ETP+ and comparison with TSEBRA

GeneMark-ETP+ achieved significantly higher prediction accuracy than the previous GeneMark versions, which integrate either RNA-Seq or protein evidence separately (Figures C.1, 5.10 and 5.11 and Table C.4). The improvements were most notable in large genomes, especially the GC-heterogeneous ones (Figure 5.11). For example, in terms of gene F1 accuracy, GeneMark-ETP+ improved over the protein-based GeneMark-EP+ by 14.1, 33.6, and 55.3 percentage points on average in the groups of compact, large homogeneous, and large heterogeneous genomes, respectively (Table C.4). The corresponding improvements with respect to GeneMark-EP+ in terms of Exon F1 accuracy were 5.2, 15.4, and 43.2; in the same groups of genomes (Table C.4). The improvements in comparison to RNA-Seq-based GeneMark-ET were even higher in all tested scenarios.

GeneMark-ETP+ also compared favorably against TSEBRA, a tool that combines predictions of RNA-Seq-based BRAKER1 and protein-homology-based BRAKER2 (Figures C.1, 5.10 and 5.11 and Table C.5). The average differences between TSEBRA and GeneMark-ETP+ in gene F1 accuracy were -2.2, 8.3, and 39.5 percentage points in the groups of compact, large homogeneous, and large heterogeneous genomes, respectively

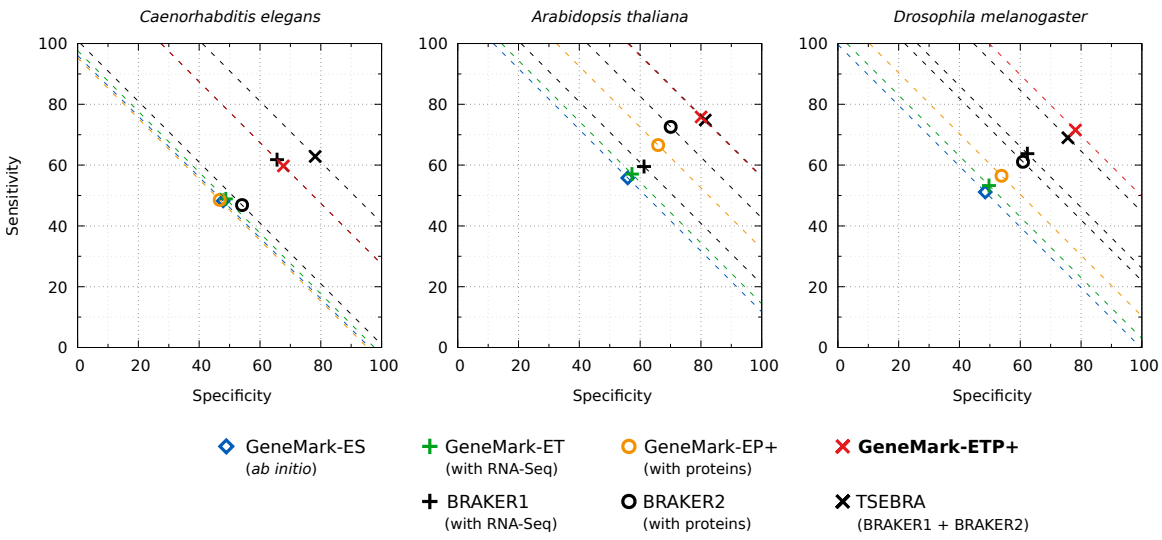

Figure 5.10: Gene level accuracy of GeneMark-ETP+ and other tools in three compact genomes. The dashed lines correspond to constant levels of $\frac{Sn+Sp}{2}$. In all tests, the protein database did not include proteins of species of the same taxonomic order as the species in question.

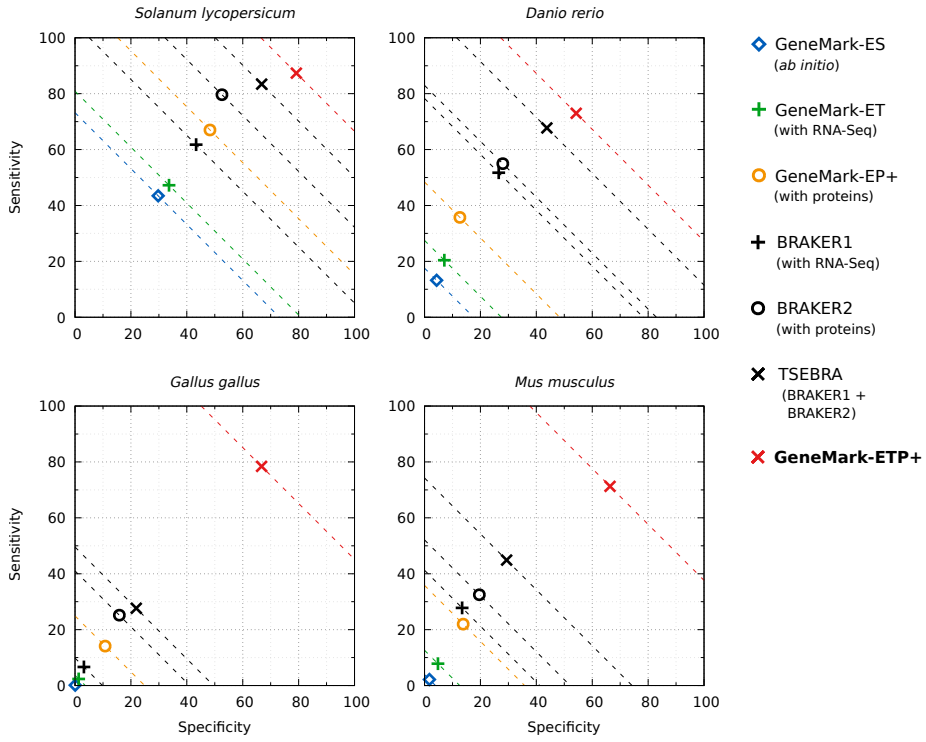

Figure 5.11: Gene level accuracy of GeneMark-ETP+ and other tools. The comparisons are the same as in Figure 5.10, but for the set of large genomes.

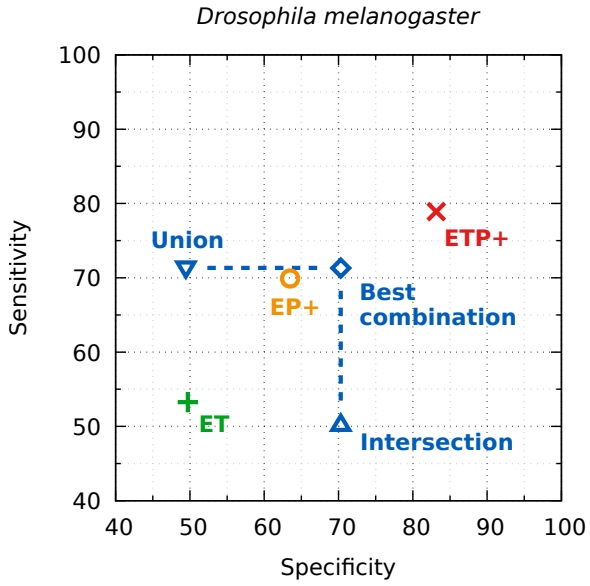

Figure 5.12: Gene-level accuracy of the optimal combination of GeneMark-ET (ET) and GeneMark-EP+ (EP+) compared to the prediction accuracy of GeneMark-ETP+ (ETP+). The result is shown for *D. melanogaster* and closely related proteins on input (only proteins of the tested species itself were excluded from the database).

(Table C.5). The improvements in terms of exon F1 values were 0.6, 1.0, and 19.1, respectively (Table C.5).

5.4.2 Results of the optimal combination of GeneMark-ET and GeneMark-EP+

The accuracy of the optimal combination of GeneMark-ET and -EP+ (see Section 5.3.5.3) was significantly lower than the accuracy of GeneMark-ETP+. This result was observed for all genomes: compact (Figure 5.12), and especially the larger genomes (Figure C.2).

5.4.3 Accuracy of the GMS-T gene prediction refinements

In GeneMark-ETP+, the protein-guided refinement of GMS-T gene predictions produces a set of high-confidence genes (Section 5.3.2). The evaluation of this procedure (Tables C.6 and 5.2) proved its importance. For both types of protein databases, with closely related (Table C.6) but also with remote proteins (Table 5.2), the refinement led to an average increase in gene-level specificity by 25 percentage points. As a result, the specificity of the

Table 5.2: Gene level accuracy of raw GMS-T predictions and the final high-confidence (HC) genes. The first column (Raw GMS-T) shows the accuracy of initial GMS-T gene predictions in all assembled transcripts. The second column (HC genes) shows the accuracy of the refined and selected HC genes. Remote proteins (proteins from species of the same taxonomic order removed from the database) were used in each case.

		Raw GMS-T	HC genes
C. elegans	Sn	**47.6**	35.8
	Sp	63.8	**88.4**
A. thaliana	Sn	51.7	**57.0**
	Sp	80.0	**97.3**
D. melanogaster	Sn	**60.5**	55.2
	Sp	82.0	**94.7**
S. lycopersicum	Sn	68.0	**75.1**
	Sp	74.6	**92.8**
D. rerio	Sn	60.5	**67.3**
	Sp	57.2	**84.6**
G. gallus	Sn	49.8	**74.7**
	Sp	43.3	**85.6**
M. musculus	Sn	50.0	**63.7**
	Sp	59.5	**90.4**

predicted set of high-confidence genes averaged over 90% on the gene level.

Furthermore, the results presented in Table 5.2 showed a significant increase in gene prediction sensitivity in five of the seven tested genomes (notably, in all seven genomes when protein database included all species other than the species of interest, see Table C.6). This increase was driven by the modifications to GMS-T predictions, especially by the improved classification of complete and incomplete GMS-T predictions (described in Section 5.3.2.2). For instance, considering transcripts with correctly predicted stop codons and no assembly errors, 2,753 (out of 22,979) predictions were classified as incomplete by GMS-T in *D. rerio*. According to the reference annotation, 1,384 of those predictions were truly incomplete while 1,369 were complete predictions misclassified as incomplete. The classification process (results in Figure 5.13) correctly identified and shortened 1,159 of the 1,369 misclassified predictions (85% sensitivity). This came at a cost of incorrectly interpreting as complete 122 genes from the group of 1,384 truly incomplete genes (9% error rate). Table C.7 shows the results of this analysis for all the tested genomes.

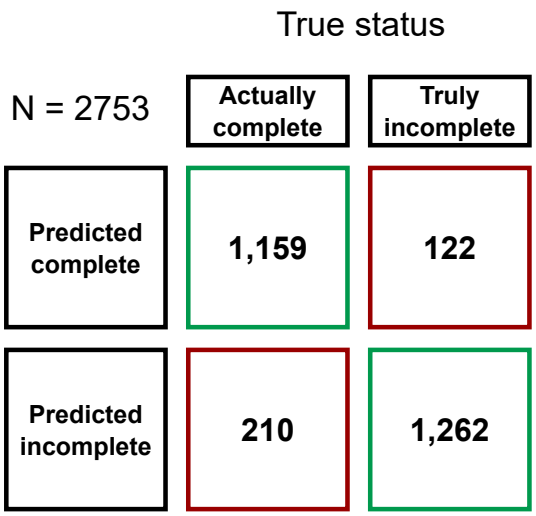

Figure 5.13: Confusion matrix for the procedure of complete/incomplete classification of GMS-T predictions (described in Section 5.3.2.2) in *D. rerio*. Remote proteins (proteins from species of the same taxonomic order removed from the database) were used on input.

5.4.4 Assessment of the repeat masking penalty estimation

The results of the experiment described in Section 5.3.5.1 showed that the automatically estimated penalty values were close to the optimal ones; i.e., to the values maximizing the gene-level F1 accuracy (Figure 5.14A). Furthermore, Figure 5.14B shows that the changes in the fraction of correctly predicted HC exons during penalty estimation reflected the changes in the GeneMark-ETP+'s overall prediction sensitivity.

The final penalty estimates for all genomes are shown in Table C.9. Notably, in GC-heterogeneous genomes, GeneMark-ETP+ estimates an optimal masking penalty for each of the GC bins; Table C.9 thus also shows how the estimated penalty values varied between GC bins. Finally, the amount of repeat-masked sequence (masked by RepeatModeler2/RepeatMasker) in each genome is shown in Table C.10.

5.4.5 Effect of the filtering of pure *ab initio* predictions

The filtering of pure *ab initio* predictions (described in Section 5.3.4.2) ensures high prediction specificity in large genomes. The evaluation of this procedure (Table C.8) showed

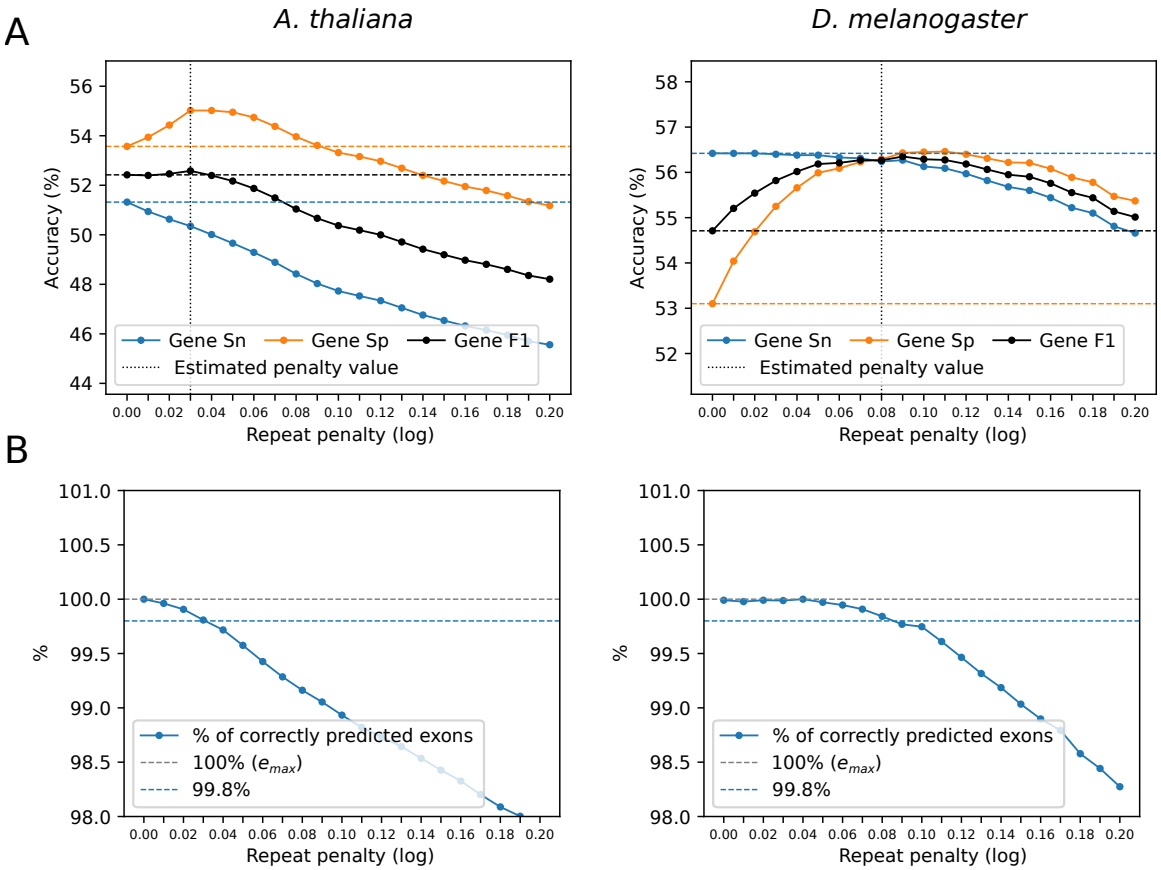

Figure 5.14: (A) The dependence of the final gene prediction accuracy on the repeat masking penalty values. The gene prediction accuracy was computed against the reference annotation, but only in the non-HC-segments as HC genes themselves are not affected by the repeat penalty. (B) The dependence of the % of correctly predicted HC exons during penalty estimation (see Section 5.3.3.4) on the masking penalty.

the following trends. In the three compact genomes ($<$ 300 Mbp), the filtering, on average, increased the gene-level specificity by 3.4 percentage points at the cost of a 2.5 sensitivity decrease. Conversely, in the four large genomes ($\geq$ 300 Mbp), the filtering, on average, increased the gene-level Sp by 19.6 percentage points at the cost of a 0.4 Sn decrease.

5.5 Discussion

The main reason to develop GeneMark-ETP+ was a clear need to create a self-training algorithm that would utilize the full extent of information in diverse information streams—genomic, transcriptomic, and cross-species proteins. The design of GeneMark-ETP+ was largely motivated by our attempts to combine RNA-Seq and protein-homology evidence in the prediction steps of existing GeneMark-ET and -EP+, since such an approach produced only marginally better results than GeneMark-EP+ alone. Thus, we had to develop a new algorithm that would simultaneously utilize all the information streams throughout all stages of its model training and gene prediction.

In this section, we describe specific factors contributing to GeneMark-ETP+'s high prediction accuracy, other than the integration of available evidence sources in all training and prediction steps. Next, we discuss how GeneMark-ETP+ adapts to variations in the size and quality of the input transcript and protein data. We also discuss the main design decisions behind the algorithm for the refinement of GeneMarkS-T predictions, which sits at the core of GeneMark-ETP+. Finally, we compare GeneMark-ETP+ with previous GeneMark versions and with TSEBRA, a tool that combines transcriptome- and protein homology-based results of BRAKER1 and BRAKER2.

5.5.1 Sources of accuracy improvement

5.5.1.1 Refinement of GMS-T predictions and high-confidence genes

Transcriptomic assemblies reconstructed from short RNA reads are highly unreliable [44]. The task of protein-coding gene prediction in transcripts is further complicated by the pres-

ence of numerous long non-coding RNAs. Finally, computational gene predictions (such as the ones made by GMS-T) can be wrong even in correctly reconstructed transcripts, especially when the length of the assembled 5' UTR sequence is short (Figure 5.3).

To avoid the transfer of assembly and GMS-T errors into final gene predictions, GeneMark-ETP+ uses protein homology to filter out and in some cases adjust wrong predictions in the transcripts. The resulting set, so-called high-confidence genes, is significantly more accurate than the initial set of all GMS-T predictions (Tables C.6 and 5.2). Thus, the algorithm for refining GMS-T predictions and selecting high-confidence genes is an essential step of GeneMark-ETP+. Without this step, which we believe is unique to GeneMark-ETP+, the raw GMS-T predictions in all transcript assemblies would not be accurate enough to be used either (i) as a training set, or (ii) directly mapped to genomic DNA to constitute the final predictions in corresponding genomic loci.

5.5.1.2 GC-specific training and predictions

The addition of GC-specific training (both in GMS-T and genomic GHMM) and predictions is a critical new feature because the absence of multiple GC models prevented previous GeneMark versions from reaching high accuracy in GC-heterogeneous species, such as the mammalian genomes (as discussed in Section 3.5.4.3). This explains why the difference between GeneMark-ETP+ and the older GeneMark versions was by far the highest in GC-heterogeneous genomes (Figure 5.11).

Still, GeneMark-ETP+ has a "GC-homogeneous" mode, which is applied to genomes in which GC-heterogeneity was not detected (Section 5.3.3.1). We observed that in GC-homogeneous genomes, using the GC-heterogeneous path led to an increased runtime and a slight accuracy decrease. We speculate that the decrease in accuracy was caused by splitting the training set into three smaller subsets; thus making the training less stable. For these reasons, GeneMark-ETP+ automatically detects GC-heterogeneity to decide whether to apply the GC-specific training and predictions.

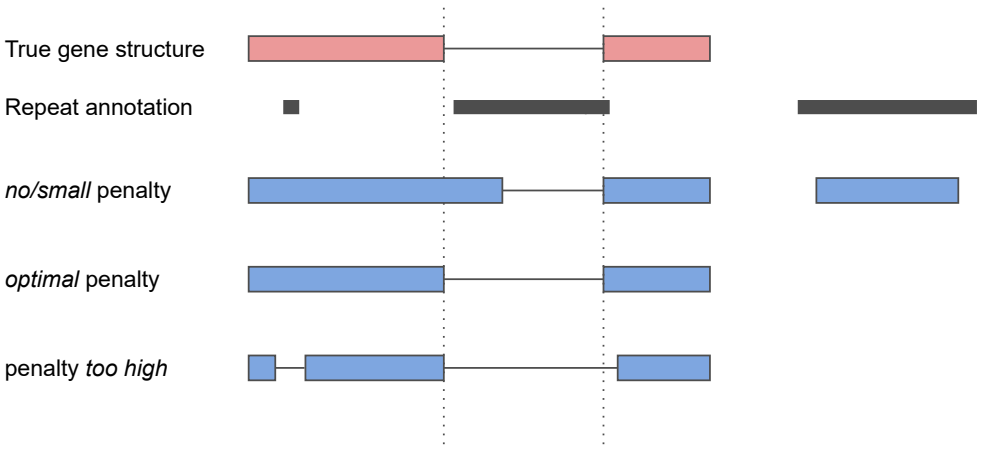

Figure 5.15: The importance of choosing the optimal repeat penalty.

5.5.1.3 Integration of repeat annotation evidence

Repetitive elements make gene prediction difficult because they often contain open reading frames with a composition similar to actual coding genes. This causes portions of transposons to be predicted as exons, thus ruining the gene predictions [3, 124]. To mitigate this issue, most gene prediction algorithms incorporate the annotation of repeats into the gene prediction process. Unfortunately, repeat annotations are usually imperfect because computational repeat masking methods are prone to masking parts of true coding genes with conserved functional motifs [125]. Therefore, the annotated repeats cannot be simply excluded (hard-masked) from the sequence.

The previous versions of GeneMark incorporated repeat annotation by hard-masking all repeats longer than a threshold T and ignoring the rest of the repeats (by default, $T = 1000$ nt for genomes shorter than 300 Mbp, and $T = 100$ nt for longer genomes). Such an approach was not optimal because (i) the information in repeats shorter than T is lost and (ii) any coding exons overlapped by repeats longer than T cannot be predicted correctly. GeneMark-ETP+'s probabilistic repeat-integration approach, in which the coding probability is decreased proportionally to the length of the repeat overlap, solves both of these issues.

To our knowledge, most other gene prediction algorithms and pipelines adopt the hard-masking strategy or attempt to filter out repeat-containing predictions in a post-processing step. AUGUSTUS [12, 105–108] is an exception as it uses a probabilistic repeat integration. In fact, the repeat masking integration of ETP+ was directly inspired by this approach. Still, AUGUSTUS uses a fixed default repeat penalty value for all input genomes. Our results (Figure 5.14 and Table C.9) suggest that the optimal value is specific to each input genome; Figure 5.15 illustrates the importance of choosing the optimal value. The ability to automatically estimate this species-specific penalty value (Figure 5.14) contributes to the overall increase in GeneMark-ETP+'s prediction accuracy. While the repeat penalty estimates varied between different genomes, they were stable with respect to changing sets of input proteins (Table C.9), and thus to the number of high-confidence genes used for the penalty estimation.

5.5.1.4 Filtering of pure ab initio predictions

The filtering of pure *ab initio* predictions in non-HC-segments (Section 5.3.4.2) is critical to ensure high prediction sensitivity in long genomes with a large intergenic space—the filtering increased the prediction specificity by $\sim$20 percentage points in this group of genomes (Table C.8). To a lesser extent, the filtering also improved the prediction specificity in compact genomes (Table C.8). However, in these smaller genomes, the specificity increase came at a cost of a comparable sensitivity decrease; thus it is not applied in genomes $<$ 300 Mbp.

These results may create an impression that the *ab initio* component does not play any useful role in large genomes. However, it is important to realize that only exclusively *ab initio* predictions, with no protein or transcript support, are removed. A significant portion of predictions may be only partially supported by the extrinsic evidence and the *ab initio* component is essential to connect the partial evidence into full gene structures.

5.5.2 Adapting to variations in the size of extrinsic inputs

The main focus of GeneMark-ETP+ is on the optimal integration of RNA-Seq and protein evidence sets. While optimal prediction accuracy is achieved when both of these inputs are large (i.e., sufficient transcriptome coverage and closely related proteins), we designed GeneMark-ETP+ to also work well when one of these inputs lacks in size.

One such design decision was to train the genomic GHMM model in a semi-supervised way (as in GeneMark-ET, and -EP+) when the set of high-confidence contains less than 4,000 genes. This way, a high-quality genomic model, which can capture the genes missing in the high-confidence set, is trained even when the extrinsic input set is limited. We observed that the use of the extended training did not improve prediction accuracy when $>$ 4,000 training genes were available.

Another decision, aimed at cases when the input proteins are only remotely related to the genome of interest, was to add GMS-T predictions with no protein but strong intrinsic support into the set of high-confidence genes (Section 5.3.2.3). As a result, compared to GeneMark-EP+ (which uses proteins only), GeneMark-ETP+ is less affected by the changes in the input protein database. For example, the accuracy of GeneMark-EP+ decreased by 11.4 percentage points (gene-level F1) when remote proteins were used in place of closely related ones in *D. melanogaster* (Table C.4). In the same situation, the accuracy of GeneMark-ETP+ decreased by 6.3 percentage points. Here it is important to note that proteins are still utilized to improve the quality of high-confidence genes with intrinsic support because predictions in conflict with input proteins are removed from the high-confidence set (Sections C.1 and 5.3.2.3). This proved to be an important filtering step that removes false predictions caused by partially incorrect assemblies (e.g., with an in-frame intron retention).

5.5.3 Design decisions for the refinement of GMS-T predictions

5.5.3.1 Removal of 3' incomplete GMS-T predictions

As described in Section 5.3.2.2, GeneMark-ETP+ attempts to improve the accuracy of GMS-T predictions that were predicted as 5' incomplete. The GMS-T predictions can also be incomplete at their 3' ends (unambiguously defined by the lack of a stop codon). We observed that 3' incomplete predictions were much less common than their 5' counterparts and that they usually indicated a prediction error rather than an actual 3' incomplete assembly. This was not unexpected since most RNA-Seq libraries are prepared using the poly-A tail enrichment of mRNA transcripts and are thus biased towards coverage on 3' ends [130]. For these reasons, all 3' incomplete predictions are removed from the high-confidence gene set.

5.5.3.2 Reliability of complete GMS-T predictions

GMS-T often incorrectly predicts that a truly complete transcript with a short 5' UTR is incomplete (Figure 5.4); GeneMark-ETP+ attempts to fix these errors (Section 5.3.2.2). The corresponding type of a GMS-T error—incomplete assemblies misclassified as complete coding—can also occur, but this is much less common in the GMS-T output. We observed that most genes predicted as complete by GMS-T have an upstream in-frame stop codon in the 5' UTR and/or a strong translation initiation signal (TIS). For the remaining predictions, with a weak TIS and no in-frame stop codon in the 5' UTR, we attempted to use the scoring system from Equation (5.1) to improve the classification of complete and incomplete predictions. However, in our experiments, this classification was not able to significantly improve the original (mostly correct) GMS-T labeling, and it is therefore not used in the algorithm.

5.5.3.3 *Adjustment of predictions creating less than longest ORF*

Most transcripts are translated from the start codon closest to the 5' end [131]. Still, the translation can be initiated at a downstream start; e.g., when the upstream starts have a weak translation initiation signal (the Kozak pattern [132]). GMS-T accounts for the possibility of non-5'-most translation starts by predicting the translation start based on the strength of its Kozak pattern (derived in species-specific self-training). Because the Kozak pattern is a relatively weak signal (and possibly because of the bias in reference annotations towards the 5'-most translation initiation sites), the non-5'-most GMS-T start predictions exhibit a higher false-positive rate when compared to their 5'-most start counterparts. To mitigate this issue, GeneMark-ETP+ extends predictions with the non-5'-most start to the longest open reading frame when the longer prediction is well-supported by external protein evidence (Section 5.3.2.4). As a result, the high-confidence gene set can still contain non-5'-most predictions, but their error rate is significantly lower than the original set of all such GMS-T candidates.

5.5.3.4 *Derivation of classification scores*

To derive Equations (5.1) and (5.2), we trained random forest and logistic regression classifiers—using all alignment features offered by DIAMOND's tabular output—to classify predictions as complete/incomplete (for Equation (5.1)), or true/false (for Equation (5.2)). The ground-truth labels were determined by comparisons with reference annotations; the set of training data contained representative GMS-T predictions from each group of genomes (Table 5.1). Next, we explored the trained models to determine which alignment features were the most important for correct classification. Finally, by trying multiple designs, we found how to best combine the most important features into single scores with a straightforward biological interpretation, giving rise to Equations (5.1) and (5.2). Notably, when used to classify GMS-T predictions in a hold-out test set, the application of Equations (5.1) and (5.2) led to better prediction accuracy than the classification with the trained random

forest and logistic regression models.

5.5.4 Comparison of GeneMark-ETP+ with other gene finders

5.5.4.1 Comparison with GeneMark-ET, -EP+, and their optimal combination

In all tested genomes, GeneMark-ETP+ significantly outperformed the previous GeneMark versions—the RNA-Seq-based GeneMark-ET and protein-homology-based GeneMark-EP+. This improvement may not be surprising as GeneMark-ET and -EP+ each use only a single source of extrinsic evidence. Therefore, we also compared the accuracy of GeneMark-ETP+ to the accuracy corresponding to an ideal combination of gene sets predicted by GeneMark-ET and -EP+ (Section 5.3.5.3). This optimal combination would have the sensitivity and specificity corresponding to the union and the intersection of gene predictions by the two tools, respectively. Assuming this ideal combination could be achieved, its prediction accuracy would still be far below that of GeneMark-ETP+ (Figures C.2 and 5.12). This comparison demonstrates that the high prediction accuracy of GeneMark-ETP+ is a result of integration of transcriptomic and protein data in *all* stages of the algorithm, and cannot be achieved in a "post-processing" step.

5.5.4.2 Comparison with TSEBRA

RNA-Seq-based BRAKER1 and protein-homology-based BRAKER2 were demonstrated to be one of the most accurate gene prediction pipelines ([47], Chapter 4). In turn, TSEBRA, a tool that finds an optimal combination of predictions made by BRAKER1 and BRAKER2, was shown to achieve higher accuracy than (i) either BRAKER1 or BRAKER2 alone, and (ii) EVidenceModeler [76], one of the most prominent combiner tools. Therefore, TSEBRA is an excellent representative of tools combining RNA-Seq- and protein homology-based predictions.

In the group of large genomes, GeneMark-ETP+ achieved significantly higher accuracy than TSEBRA (Figure 5.11). This was especially true for the GC-heterogeneous genomes

(*G. gallus*, *M. musculus*), which is not surprising because BRAKER1 and BRAKER2 do not adjust for variations in GC content (as discussed in Section 4.5.1.2). However, GC-heterogeneity is not the only reason for GeneMark-ETP+'s improved accuracy in large genomes, as it showed significant improvements over TSEBRA in the GC-homogeneous genomes of *S. lycopersicum* and *D. rerio* as well. The accuracy differences were much smaller in the group of compact genomes (Figure 5.10), with TSEBRA achieving higher accuracy in *C. elegans*.

The fact that neither tool was clearly better in all tested genomes is actually encouraging, as it hints at the possibility of joining the two algorithms. In the spirit of BRAKER1 and BRAKER2, a gene prediction pipeline, BRAKER3, could be developed to combine the strengths of GeneMark-ETP+, AUGUSTUS, and TSEBRA.

5.6 Conclusion

GeneMark-ETP+ is a self-training eukaryotic gene prediction algorithm that combines genomic, transcriptomic, and protein homology information sources throughout all stages of its automatic model training and gene prediction. We observed that GeneMark-ETP+ delivered high prediction accuracy in all tested genomes, including the group of difficult-to-annotate, large, GC-heterogeneous genomes. Furthermore, GeneMark-ETP+ achieved significantly better prediction accuracy than any combination of previous GeneMark versions that use either transcriptomic or protein homology evidence.

5.7 Availability

We are currently finalizing the GeneMark-ETP+ distribution package. GeneMark-ETP+ will be available on GitHub and at http://topaz.gatech.edu/GeneMark/license_download.cgi. The runtime of GeneMark-ETP+ scales with the genome size and is comparable to the one of GeneMark-EP+. For example, on a machine with 64 CPU cores, with the genomes of *D. melanogaster*, *D. rerio*, and *M. musculus* on input, the runtimes were 1, 4.5, and 6.5

hours, respectively.

CHAPTER 6
CONCLUSION

This dissertation presented three novel algorithms for automatic gene prediction in eukaryotic genomes, each of which solved several problems that hindered the accuracy and usability of existing gene prediction methods.

First, we presented GeneMark-EP+, an unsupervised gene prediction algorithm that uses homologous cross-species proteins to guide its model training and gene prediction steps. The main reason to develop GeneMark-EP+ was a clear need to leverage abundant protein sequence data available in public databases for improving the accuracy of automatic gene prediction. The use of protein homology in gene prediction poses a challenge due to the patchiness of the evidence proteins generate and the decrease in prediction accuracy with the increasing evolutionary distance of proteins. GeneMark-EP+ addressed this challenge by finding an optimal method of homologous protein evidence incorporation into the automatic iterative training of an *ab initio* algorithm. The need to process large protein databases and determine the reliability of the mapped protein evidence led to the development of a new pipeline called ProtHint—a tool that predicts accurate locations of exon boundaries from a large number of proteins of *any evolutionary distance* to the genome of interest. The novel scoring system developed in ProtHint made it possible to define protein hints with over 95% specificity regardless of the number and evolutionary distance of target proteins. Due to their high specificity, these reliable hints could be directly incorporated into the final GeneMark-EP+ predictions; thus significantly increasing its prediction accuracy. We showed that GeneMark-EP+ delivered better prediction accuracy than *ab initio* GeneMark-ES and RNA-Seq-based GeneMark-ET, even in situations when only evolutionarily remote proteins were used on input. GeneMark-EP+ should thus become a universal extension of GeneMark-ES because the protein databases are (unlike transcriptomic data)

always readily available prior to a genome annotation project start. Indeed, since its release in May 2020, GeneMark-EP+ has been downloaded $>$ 5000 times and references to GeneMark-EP+ (cited 80 times) have appeared in a number of genome projects annotating fungi, protists, plants, and animals.

Second, we introduced BRAKER2, a fully automated protein homology-based gene prediction pipeline that integrates ProtHint and GeneMark-EP+ with AUGUSTUS. By combining complementary strengths of multiple gene prediction tools, BRAKER2 achieves state-of-the-art gene prediction accuracy in a fully unsupervised manner. There are several reasons why BRAKER2 performs better than AUGUSTUS or GeneMark-EP+ alone. In contrast to GeneMark-EP+, AUGUSTUS allows for more flexible integration of protein hints into an *ab initio* gene prediction. This makes it possible to integrate *all* of the protein hints generated by ProtHint into AUGUSTUS predictions, not just the subset of reliable hints utilized by GeneMark-EP+. On top of that, unlike GeneMark-EP+, AUGUSTUS predicts alternative isoforms of protein-coding genes. That said, AUGUSTUS, a supervised algorithm, cannot be used at all without a reliable training set—prepared by GeneMark-EP+ in an unsupervised manner. Further, although AUGUSTUS contains a sophisticated mechanism for the integration of protein hints, the task of the actual preparation and scoring of such hints is solved by ProtHint. We demonstrated that BRAKER2 achieves high prediction accuracy in the absence of closely related proteins, and illustrated how this is facilitated by an optimal AUGUSTUS training and the simultaneous use of multiple reference proteins. We showed that even in tests with the most remotely related proteins, the accuracy of BRAKER2 was comparable to that of BRAKER1, which was supported by a large amount of RNA-Seq data. Finally, we showed that BRAKER2 achieved significantly higher prediction accuracy than MAKER2, one of the most frequently used gene prediction pipelines. The usefulness of BRAKER2 has been verified by its use in numerous genome annotation projects; since its release in January 2021, BRAKER2 has been referenced by close to 200 publications.

Finally, we presented GeneMark-ETP+, a self-training eukaryotic gene prediction algorithm that combines genomic, transcriptomic, and protein homology information sources throughout all stages of its automatic model training and gene prediction. The main reason to develop GeneMark-ETP+ was a clear need to create a self-training algorithm that would utilize the full extent of information in the mentioned information streams. GeneMark-ETP+ facilitates the evidence integration by, among other things, creating a novel method for simultaneous gene prediction in transcripts (assembled from RNA-Seq) and genomic DNA. Importantly, the training of GeneMark-ETP+ is fully unsupervised and the protein homology evidence integration utilizes proteins of any evolutionary distance, including remote homologs; thus building upon the work described in the previous two chapters. We described the salient components of GeneMark-ETP+—a novel method that uses protein homology to refine GMS-T predictions in transcripts, GC-content specific automatic training, and a novel mechanism of integrating repeat annotations—and showed how each of these components contributed to GeneMark-ETP+'s high gene prediction accuracy. We showed that GeneMark-ETP+ achieved significantly better accuracy than any combination of previous GeneMark versions (-ET, -EP+) that use either transcriptomic or protein homology evidence. We also demonstrated that in large eukaryotic genomes, GeneMark-ETP+ was more accurate than TSEBRA, a combiner of BRAKER1 and BRAKER2 predictions. Overall, we expect GeneMark-ETP+ (manuscript in preparation) to be an important step toward fully automated and accurate eukaryotic gene prediction.

Despite the advances described in this thesis, there are still opportunities for further enhancements. For instance, as hinted at the end of Chapter 5, the strengths of GeneMark-ETP+ and AUGUSTUS could be combined to develop a gene prediction pipeline, BRAKER3. To outline BRAKER3, the high-confidence genes predicted by GeneMark-ETP+ would make an ideal set for AUGUSTUS training, including the training of GC-specific models. AUGUSTUS could, in turn, be used to predict more accurate genes in non-HC-segments, including predictions of alternative isoforms; as GeneMark-ETP+ only predicts

alternative isoforms in the high-confidence loci. Another opportunity for improving gene prediction methods lies in the utilization of long RNA reads. Although GeneMark-ETP+ was designed to directly work with transcripts assembled from long-read RNA sequencing, careful tests must be conducted to determine whether long reads positively contribute to GeneMark-ETP+'s prediction accuracy and whether the design of GeneMark-ETP+ needs to be adjusted to better account for the long reads.

Appendices

APPENDIX A

GENEMARK-EP+

A.1 Accuracy assessment of GeneMark-EP+ on exon level

The exon level accuracy of GeneMark-EP+ (Figure A.2) followed the same trends as the gene level accuracy described in Section 3.4.1.

A.1.1 Fungal genomes (*N. crassa*)

In comparison with GeneMark-ES, we observed small improvements in GeneMark-EP+ (by $\sim$2 percentage points) when the hints originated from proteins of species outside of the *N. crassa* genus and order (Figure A.2A). No difference between -ES and -EP+ was observed when the hints came from proteins outside of the *N. crassa* phylum.

A.1.2 Compact eukaryotic genomes (*C. elegans*, *A. thaliana*, and *D. melanogaster*)

GeneMark-ES was quite accurate within this group of genomes, still the prediction accuracy of GeneMark-EP+ was higher. The improvements were most pronounced for *A. thaliana* (Figure A.2C) and *D. melanogaster* (Figure A.2D). GeneMark-EP+ with hints of proteins from the relevant genus and beyond improved over GeneMark-ES by 5–10 percentage points in both Sn and Sp. This improvement was reduced when the evolutionary distance to target proteins increased. However, even for more distant target proteins, situated outside the relevant phylum, we saw an increase in specificity by 2–4 percentage points. For *C. elegans* (Figure A.2B), the accuracy of GeneMark-EP+ improved slightly over -ES when target proteins from inside the same genus were admitted but remained almost the same when target proteins were selected from species outside the *C. elegans* genus or phylum.

A.1.3 Large eukaryotic genomes (*S. lycopersicum* and *D. rerio*)

GeneMark-ES was less accurate for large genomes than for the compact genomes. When proteins of species inside the same phylum could be used as targets for hints generation, GeneMark-EP+ showed significant increases in performance (Figure A.2EF) with Sn $\sim$75% comparable to the Sn values reached for the compact genomes. The Sp value was improved to 55%–60%. Still, this was much lower than the average Sp observed for compact genomes (a part of this difference could be attributed to the quality of the reference annotations, see Section 3.4.1.3). The improvement of GeneMark-EP+ over GeneMark-ES was by $\sim$10 percentage points in Sn and Sp (Figure A.2EF). This improvement remained high even when more remote target proteins were used for hints generation, i.e., from species outside the same phylum.

A.2 Details of the ProtHint design

A.2.1 IBA score for an exon with a frameshift

If the spliced alignment contains a frameshift, we modify the protein alignment downstream from the frameshift point (for a downstream exon) or upstream from the frameshift point (for an upstream exon) by replacing each translated codon with a gap. Each such artificial gap adds a penalty $= -4$ during the computation of the IBA score (Section 3.3.2.2).

A.2.2 Comparison between intron border alignment (IBA) and intron mapping coverage (IMC) Scores

A direct comparison of Sp-Sn curves in Figure 3.2 is not entirely fair for the following reason. All introns are filtered with IBA ≥ 0.1 and AEE ≥ 25 prior to computing the IMC score (Section 3.3.2.3). This removes a significant number of false predictions (Figure A.1). Thus, the IMC score is computed from a set already filtered with the IBA score.

A.2.3 Invariance of the spliced alignment with respect to alignment tools

ProtHint also supports the use of ProSplign as an alternative to generating spliced alignments with Spaln. We observed that the accuracy of hints generated by ProSplign as well as the accuracy of subsequent GeneMark-EP+ gene predictions did not differ significantly from the results obtained with Spaln. Since Spaln is significantly faster, it is used by default. ProtHint also supports an alternative to DIAMOND, a more sensitive but slower BLASTp. We have not observed a significant difference in ProtHint accuracy when either DIAMOND or BLASTp was used. Since DIAMOND is several orders of magnitude faster than BLASTp, ProtHint uses DIAMOND by default.

A.2.3.1 Differences between the usage of ProSplign and Spaln

ProSplign has a built-in filtering procedure; therefore the initial filtering steps described in the main text can be skipped and all hints mapped by ProSplign can be used directly. Still, the procedure of scoring and selecting high-confidence hints remains the same.

The slow speed of ProSplign hampers its use. ProSplign does not use heuristics to speed-up its dynamic programming based alignment algorithm; therefore it is 10-100x slower than Spaln, depending on the length of the genome locus and the length of the protein being aligned. To run ProSplign in a reasonable time, the "ProSplign mode" of ProtHint works as follows. ProtHint first runs Spaln to generate a set of hints. For each hint mapped by Spaln, the top ten supporting proteins are selected and aligned with ProSplign. This selection reduces the number of target proteins to be aligned by ProSplign by an order of magnitude.

We observed that the raw set of hints mapped by ProSplign was generally less sensitive and more specific than hints produced by Spaln, due to ProSplign's internal filtering procedure. However, the set of high-confidence hints was almost the same for both tools, meaning that our scoring system was insensitive to the choice of a spliced alignment engine. Consequently, the results of GeneMark-EP+ did not significantly change when either

Spaln- or ProSplign-generated alignments were used. Currently, Spaln is used as the default option in ProtHint due to its superior speed.

A.2.4 Use of a Custom Protein Database

A custom protein database could be used as an alternative to OrthoDB. A special attention should be paid to the construction of such database, as the presence of identical proteins (for example, proteins from subspecies of the same species) can lead to artificially inflated coverage as well as increase in the execution time.

A.3 Supplementary Figures

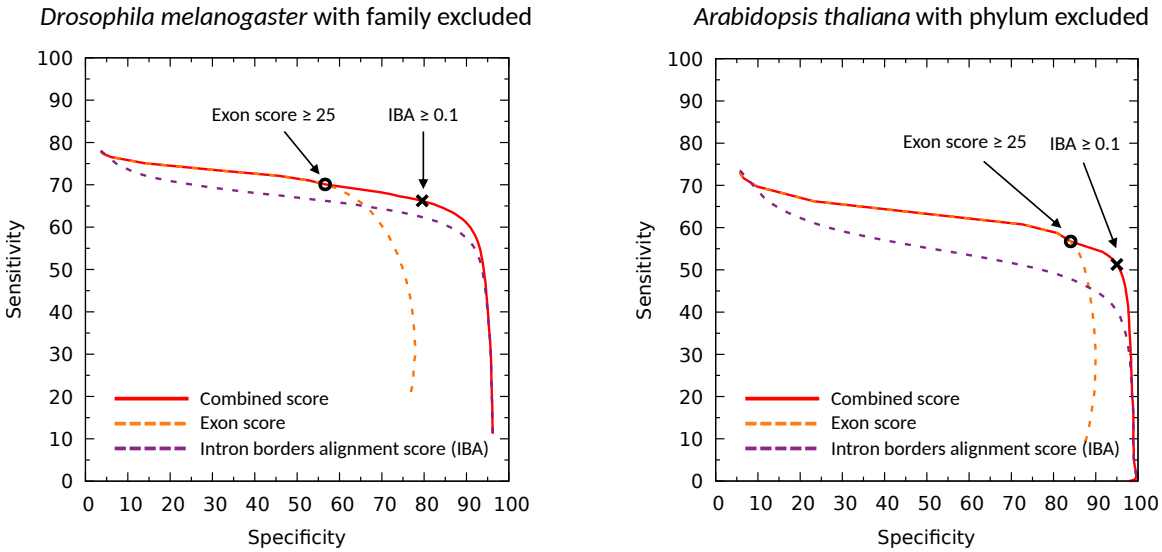

Figure A.1: ProtHint intron Sp-Sn curves built upon filtering sets of mapped introns by exon AEE scores (dashed orange) and intron borders alignment score (IBA, dashed purple). The combined curve (red) is generated by, first, selecting out all introns with AEE scores above the threshold changing from 0 to 25; next, all the selected introns are checked for having IBA scores above the threshold changing from 0 to 0.1 and up to 1.0. The position of the black cross in the combined curve represents IBA score ≥ 0.1 and AEE score ≥ 25.

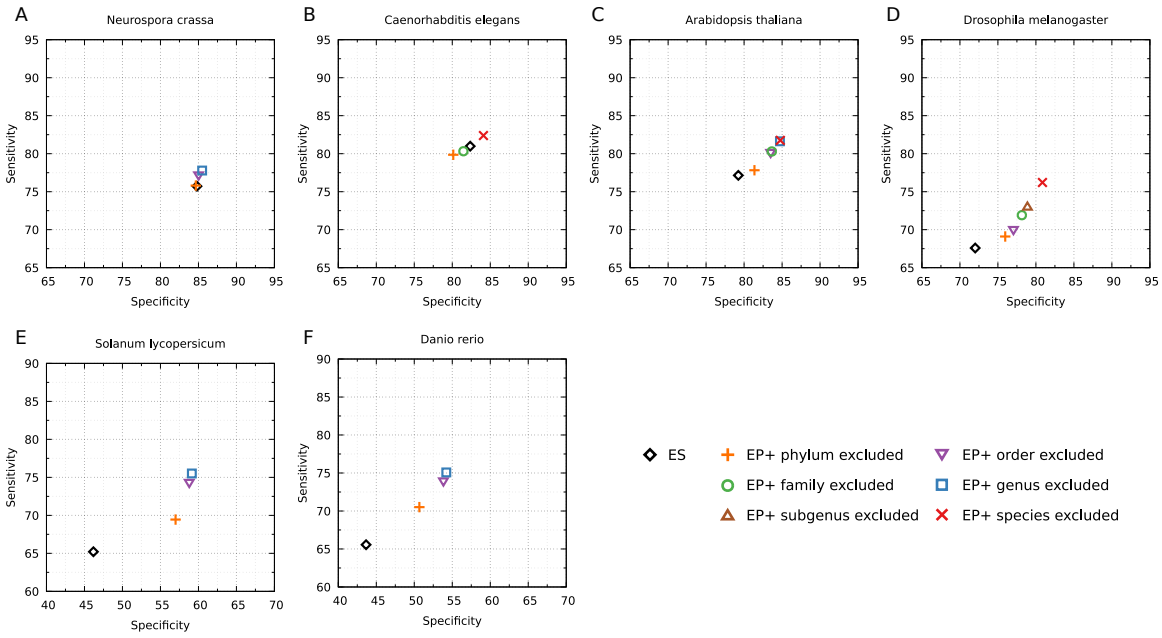

Figure A.2: A comparison of GeneMark-ES and GeneMark-EP+ accuracy on the exon level. The accuracy of GeneMark-EP+ is shown for cases when ProtHint works with different in size sets of reference OrthoDB proteins: from the largest (only the same species excluded) to the smallest (the whole same phylum excluded). Exon level Sn and Sp are defined with respect to a full complement of annotated exons, including alternative types.

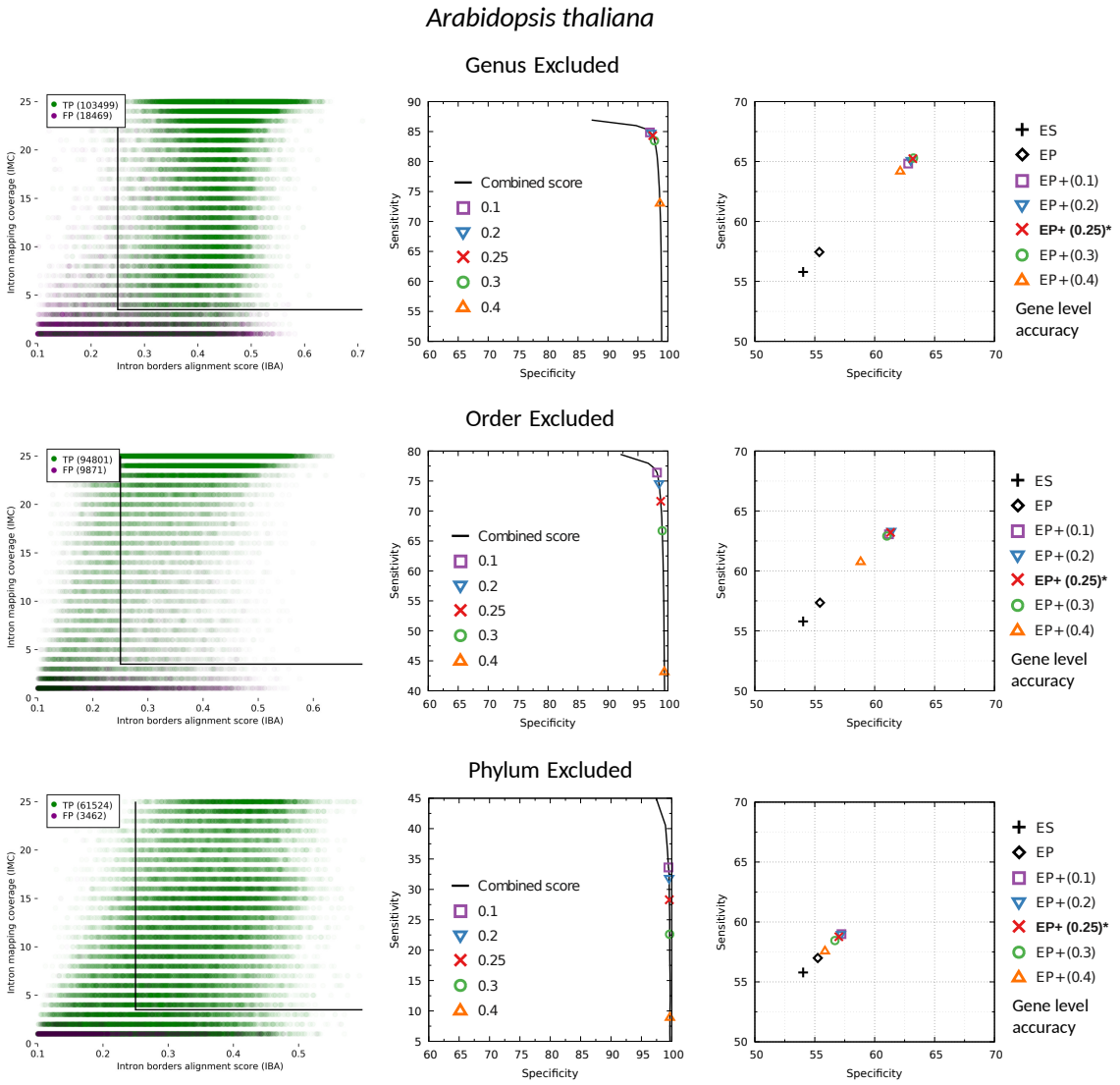

Figure A.3: The Effect of the IBA threshold on the accuracy of high-confidence hints and GeneMark-EP+ for *A. thaliana*.

Left side graphs show distributions of score vectors of true positive (green) and false positive (purple) introns (mapped and scored by ProtHint), the vectors' components are intron borders alignment (IBA) and intron mapping coverage (IMC) scores. The black lines represent cutoffs at IMC = 4 and IBA = 0.25. Total numbers of false and true positives are shown in the upper left corners.

Middle graphs display ProtHint's Sp-Sn curves. The curves are generated by first, selecting out all introns below changing the IMC threshold from 0 to 4 and then selecting out all the introns with IBA score from 0 to 0.25 and up to 1.0. The Sp-Sn values for various IBA cutoffs (0.1, 0.2, 0.25, 0.3, 0.4) are shown at the curves. The curves illustrate the procedure of selecting introns mapped with high confidence.

Right side graphs display how gene-level prediction accuracy of GeneMark-EP depends on IBA score cutoffs used to select sets of high confidence introns. Sp and Sn of GeneMark-EP, i.e., without high confidence intron enforcement, as well as for GeneMark-ES, are shown too.

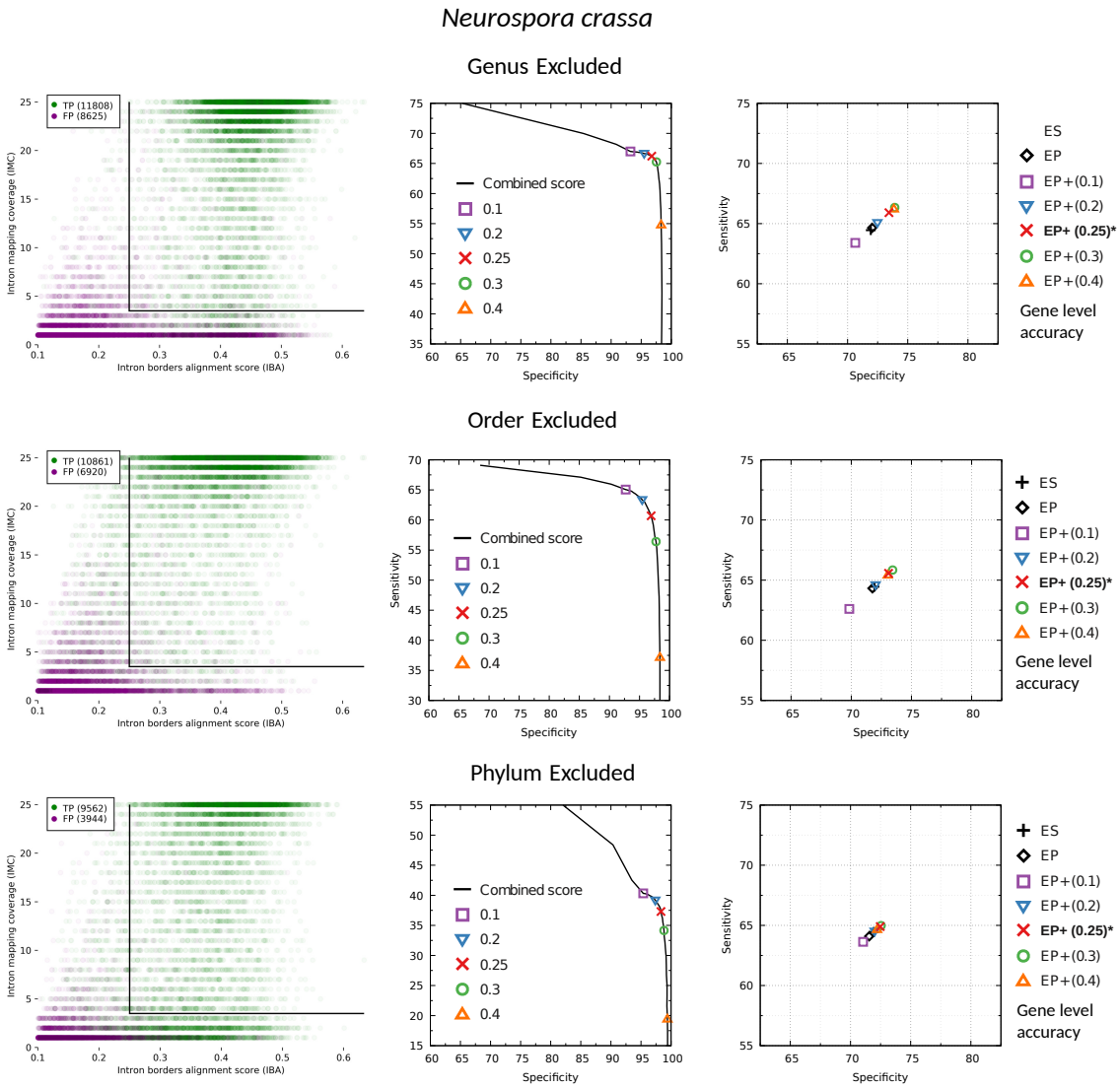

Figure A.4: The effect of the IBA threshold on the accuracy of high-confidence hints and GeneMark-EP+ for *N. crassa*. For more details see the legend to Figure A.3.

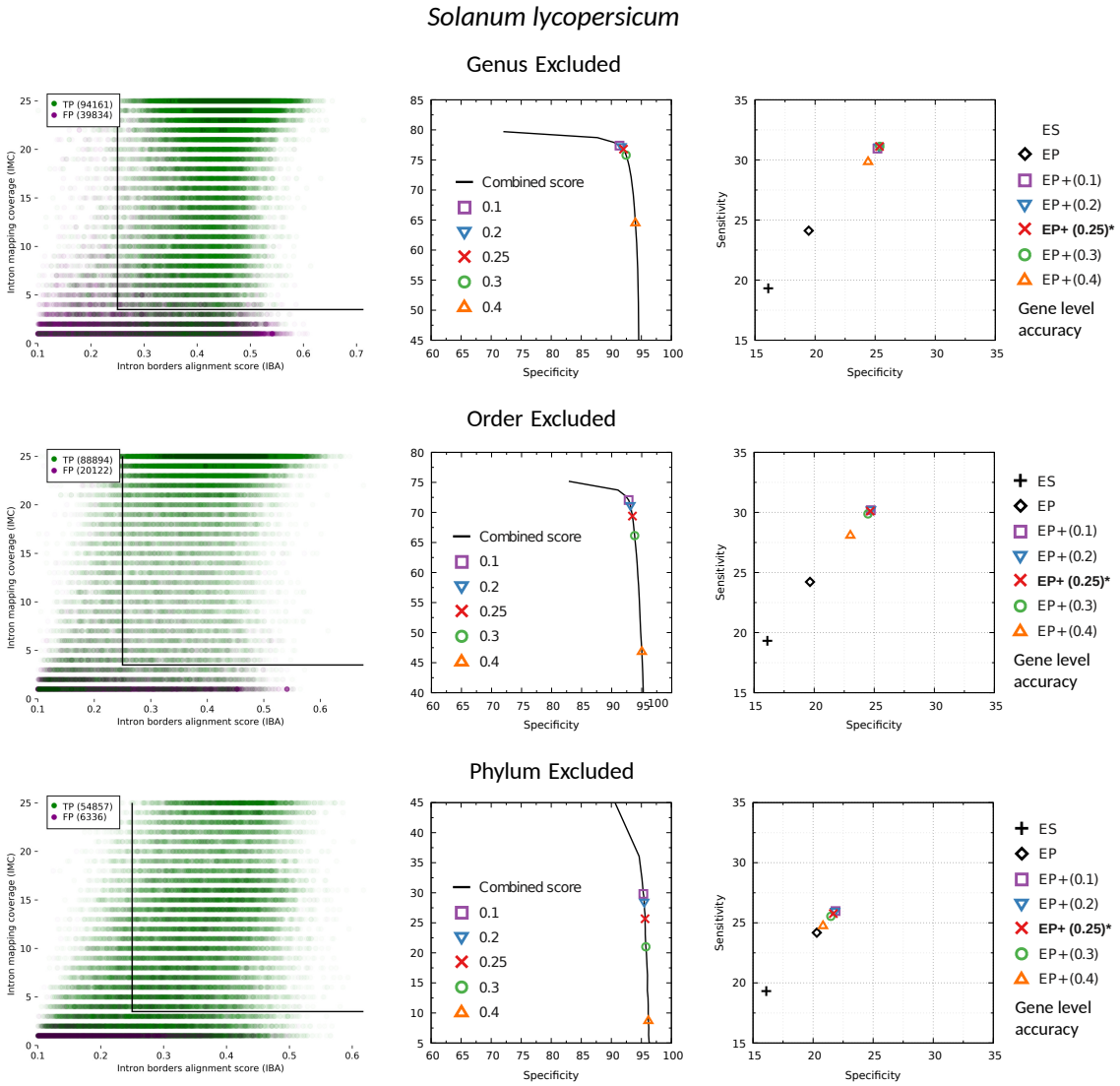

Figure A.5: The effect of the IBA threshold on the accuracy of high-confidence hints and GeneMark-EP+ for *S. lycopersicum*. For more details see the legend to Figure A.3.

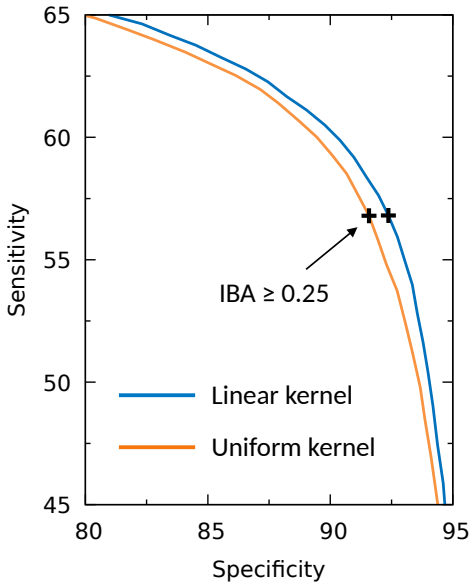

Figure A.6: ProtHint intron hint Sp-Sn curves built for intron border alignment scores (IBA) computed with the use of linear and uniform kernels (window width = 10). The crosses at the curves represent IBA score $\geq$ 0.25, with 0.25 being a value of the IBA threshold used for the high-confidence intron selection. *D. melanogaster* genome with target proteins from species outside the Drosophilidae family were used in this experiment.

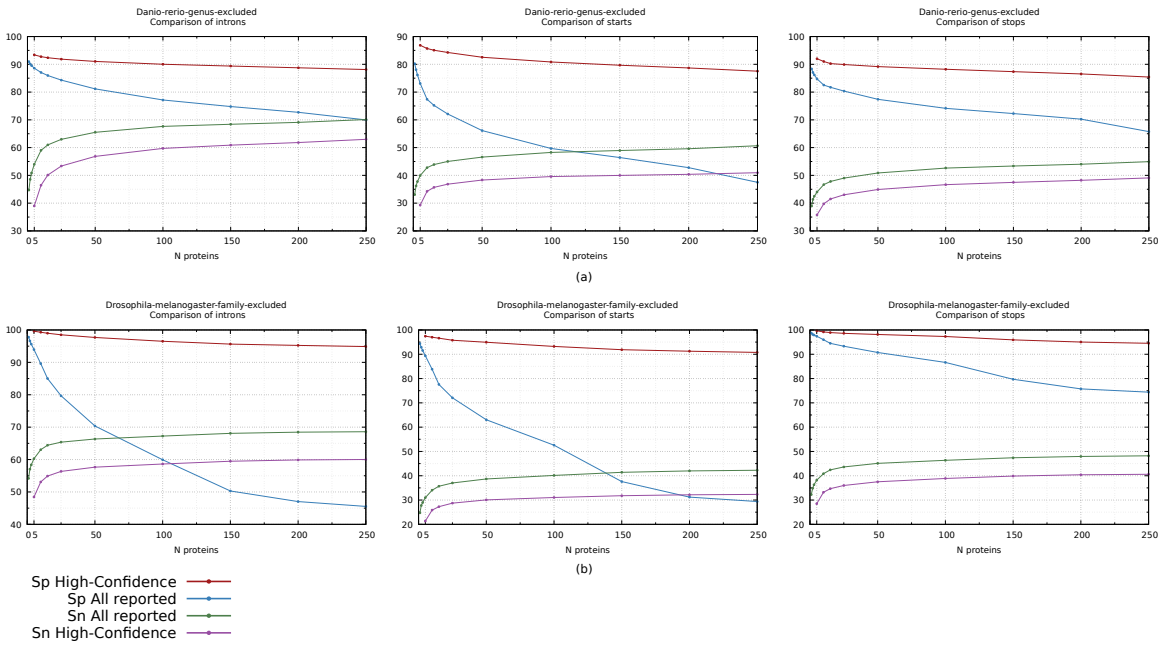

Figure A.7: The effect of the maximum number of target proteins N per seed gene on sensitivity and specificity of hints to introns, start and stop codons. All reported and high-confidence hints are shown. Number N limits how many proteins found by DIAMOND are splice-aligned back to a seed region. The examples shown are (a) for a large genome of *D. rerio*, and (b) for a compact genome of *D. melanogaster*. The increase in Sn of intron hints is larger in *D. rerio* because of a higher number of introns per gene (Table 3.1). The default value of N is set to 25 as a trade-off between computational speed of ProtHint and the Sn of produced hints. The specificity of high-confidence hints decreases slightly with the increasing N. We recommend to use more strict (higher) SMC/IMC filtering thresholds when N $>$ 25 is selected.

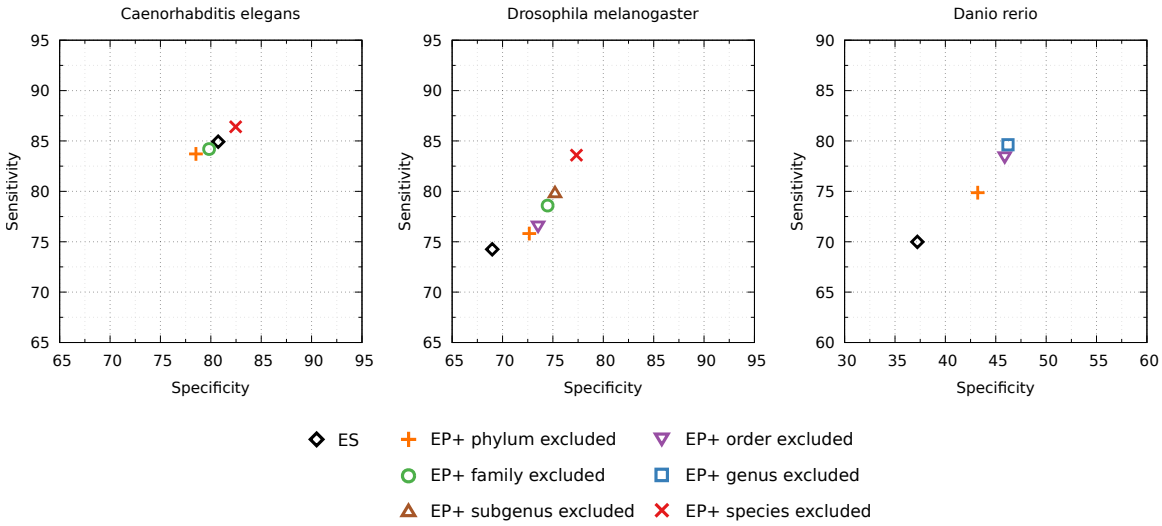

Figure A.8: The same comparison as in Figure A.2; the Sn and Sp values were computed against the APPRIS annotation of genes of gene of principal protein isoforms.

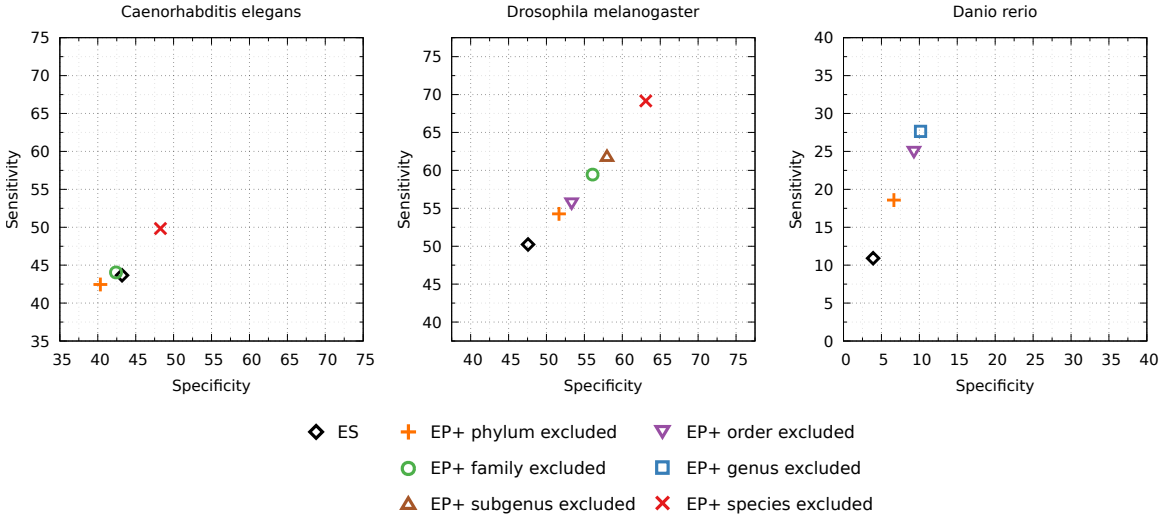

Figure A.9: The same comparison as in Figure 3.7 in the main text; the accuracy is computed against the APPRIS annotation of genes of principal protein isoforms.

A.4 Supplementary Tables

Table A.1: Sensitivity and specificity of all gene start hints created by ProtHint as well as of the high-confidence start hints. High specificity was achieved with filtering by SMC scores as well as by the removal of candidate starts overlapped by at least one target protein (suggesting that a start is located upstream). Sn was defined with respect to a full complement of starts, including alternative ones as given in annotation. The numbers were generated in tests with reference proteins from species outside the relevant genus.

		All reported starts	Filtered with SMC ≥ 4	Filtered with SMC ≥ 4 and exon overlap = 0
N. crassa	Sn	65.3	39.4	38.8
	Sp	75.6	88.5	**93.7**
C. elegans	Sn	13.4	6.1	6.0
	Sp	68.2	94.5	**95.8**
A. thaliana	Sn	69.3	62.9	61.4
	Sp	70.9	89.8	**94.4**
D. melanogaster	Sn	37.7	29.6	29.2
	Sp	71.6	92.2	**95.6**
S. lycopersicum	Sn	48.9	43.6	42.8
	Sp	39.2	65.4	**72.4**
D. rerio	Sn	47.6	40.8	39.6
	Sp	61.4	80.8	**84.1**

Table A.2: A comparison of GeneMark-ES, GeneMark-ET, GeneMark-EP and GeneMark-EP+ in terms of accuracy on gene, exon, and intron levels. Exon and intron level Sn and Sp were defined with respect to a full complement of exons/introns, including ones from alternative isoforms. The accuracy of GeneMark-EP and GeneMark-EP+ is shown for various types of protein database partition (species-excluded, etc).

			The level of exclusion of database proteins											
			Species		Subgenus		Genus		Family		Order		Phylum	
N. crassa	ES	ET					EP	**EP+**			EP	**EP+**	EP	**EP+**
Gene Sn	64.5	64.6		*		*	64.7	**67.1**		*	64.3	**66.1**	64.1	**64.7**
Gene Sp	71.9	71.9					72.0	**74.5**			71.8	**73.4**	71.6	**72.1**
Exon Sn	75.7	75.9					75.7	**77.8**			75.4	**77.2**	75.3	**75.8**
Exon Sp	84.8	84.8					85.1	**85.5**			84.9	**85.0**	84.7	**84.7**
Intron Sn	79.5	79.8					79.6	**82.5**			79.3	**82.0**	79.1	**80.3**
Intron Sp	89.8	90.0					90.5	**90.7**			90.3	**90.5**	90.0	**90.4**
C. elegans	ES	ET	EP	**EP+**					EP	**EP+**			EP	**EP+**
Gene Sn	46.8	47.8	48.7	**53.4**		*		*	45.2	**47.4**		*	43.5	**45.7**
Gene Sp	46.4	47.4	47.1	**51.8**					42.8	**45.8**			40.4	**43.6**
Exon Sn	81.0	81.0	81.3	**82.4**					80.0	**80.3**			79.6	**79.9**
Exon Sp	82.4	83.0	82.6	**84.1**					80.0	**81.5**			78.3	**80.1**
Intron Sn	87.5	87.3	87.5	**88.4**					86.4	**86.6**			86.1	**86.3**
Intron Sp	86.4	87.1	86.7	**88.1**					84.4	**85.7**			82.8	**84.5**
A. thaliana	ES	ET	EP	**EP+**			EP	**EP+**	EP	**EP+**	EP	**EP+**	EP	**EP+**
Gene Sn	55.8	57.2	57.5	**73.7**		*	57.5	**73.2**	57.3	**67.5**	57.4	**66.8**	57.0	**59.2**
Gene Sp	54.0	55.3	55.4	**69.4**			55.4	**69.1**	55.3	**64.6**	55.4	**64.0**	55.3	**57.3**
Exon Sn	77.2	77.5	77.6	**81.8**			77.5	**81.6**	77.4	**80.3**	77.5	**80.1**	77.1	**77.8**
Exon Sp	79.2	80.4	80.5	**84.8**			80.5	**84.7**	80.6	**83.7**	80.6	**83.5**	80.6	**81.4**
Intron Sn	85.2	85.5	85.5	**89.0**			85.5	**89.0**	85.4	**88.2**	85.4	**88.1**	85.1	**86.0**
Intron Sp	82.4	83.9	83.9	**87.7**			83.9	**87.7**	83.9	**87.1**	84.0	**87.0**	84.1	**85.1**
D. melanogaster	ES	ET	EP	**EP+**	EP	**EP+**			EP	**EP+**	EP	**EP+**	EP	**EP+**
Gene Sn	50.2	52.4	53.3	**69.2**	52.9	**61.8**		*	52.7	**59.5**	52.6	**55.8**	52.6	**54.3**
Gene Sp	47.6	48.8	50.0	**63.1**	49.6	**58.0**			49.6	**56.1**	50.1	**53.3**	49.7	**51.7**
Exon Sn	67.6	68.5	68.7	**76.2**	68.4	**73.0**			68.3	**71.9**	68.1	**70.0**	68.1	**69.1**
Exon Sp	72.0	73.6	74.8	**80.9**	74.5	**78.9**			74.6	**78.2**	75.1	**77.0**	74.8	**76.0**
Intron Sn	70.1	70.6	70.7	**77.6**	70.5	**75.3**			70.4	**74.3**	70.3	**72.6**	70.2	**71.6**
Intron Sp	75.5	77.3	78.7	**84.2**	78.5	**82.9**			78.6	**82.3**	79.2	**81.5**	78.8	**80.3**
S. lycopersicum	ES	ET					EP	**EP+**			EP	**EP+**	EP	**EP+**
Gene Sn	19.3	23.9		*		*	24.1	**36.3**		*	24.2	**33.5**	24.2	**26.1**
Gene Sp	16.1	19.5					19.5	**28.9**			19.7	**27.1**	20.3	**22.0**
Exon Sn	65.2	68.8					69.0	**75.5**			68.9	**74.3**	68.0	**69.5**
Exon Sp	46.2	54.0					53.7	**59.1**			54.0	**58.8**	55.7	**57.0**
Intron Sn	71.9	76.3					76.3	**84.6**			76.2	**83.5**	75.5	**77.8**
Intron Sp	48.8	59.3					58.7	**65.9**			59.1	**65.6**	61.4	**63.3**
D. rerio	ES	ET					EP	**EP+**			EP	**EP+**	EP	**EP+**
Gene Sn	12.1	16.2		*		*	16.2	**29.8**		*	16.2	**27.0**	16.4	**20.4**
Gene Sp	4.5	6.0					5.8	**11.5**			5.8	**10.6**	5.7	**7.6**
Exon Sn	64.0	66.5					66.5	**72.7**			66.3	**71.6**	66.1	**68.3**
Exon Sp	43.7	49.0					48.0	**54.2**			48.1	**53.8**	47.3	**50.7**
Intron Sn	63.7	66.2					66.3	**73.1**			66.2	**72.1**	66.0	**68.7**
Intron Sp	45.8	52.2					51.3	**58.1**			51.5	**57.8**	50.9	**54.5**

* See the first column to the right

Table A.3: A comparison of GeneMark-EP+ predictions against a full *D. rerio* annotation as well as annotation with *partial CDS* removed. Other columns show accuracy defined for a set of genes with complete/incomplete transcripts and for sets of complete/incomplete genes. A gene is considered complete if its transcripts are complete. All the numbers were generated in tests for protein database with proteins from species outside of the *D. rerio* genus.

	Original Annotation	Partial CDS removed, all transcripts	Complete transcripts	Incomplete transcripts	Complete genes	Incomplete genes
Exon Sn	69.90	72.67	**75.06**	67.60	75.08	68.71
Gene Sn	23.98	24.34	27.11	0.19	**29.84**	12.11

Table A.4: Performance of ProtHint: Sensitivity and specificity of hints to introns, gene start and stop codons. Some cells of the table are left empty due to a low number or even complete absence of species within particular taxonomic ranks (Table 3.2). The results are shown for *all reported* hints as well as *high-confidence* hints.

	The level of exclusion of database proteins											
	Species		Subgenus		Genus		Family		Order		Phylum	
N. crassa					All reported	High conf.			All reported	High conf.	All reported	High conf.
Intron Sn		*		*	76.0	66.2		*	69.9	60.7	60.2	37.3
Intron Sp					58.4	**96.8**			61.6	**96.9**	70.8	**98.3**
Start Sn					65.3	38.8			43.0	34.3	27.4	9.6
Start Sp					75.6	**93.7**			76.0	**91.5**	73.7	**89.0**
Stop Sn					65.7	40.0			44.1	35.9	29.6	10.9
Stop Sp					94.1	**98.5**			95.9	**98.4**	96.1	**99.2**
C. elegans	All reported	High conf.					All reported	High conf.			All reported	High conf.
Intron Sn	76.7	36.7		*		*	37.4	18.1		*	26.0	12.9
Intron Sp	91.8	**99.0**					92.8	**99.3**			93.7	**99.2**
Start Sn	47.7	13.4					13.4	6.0			8.2	5.1
Start Sp	75.8	**96.5**					68.2	**95.8**			76.2	**95.0**
Stop Sn	54.8	18.1					18.9	8.8			10.8	7.3
Stop Sp	90.7	**97.0**					92.4	**97.7**			92.9	**97.3**
A. thaliana	All reported	High conf.			All reported	High conf.	All reported	High conf.	All reported	High conf.	All reported	High conf.
Intron Sn	88.4	85.0		*	87.9	84.3	82.6	74.2	80.3	71.6	51.3	28.3
Intron Sp	85.8	**97.3**			86.0	**97.5**	90.9	**98.8**	91.2	**98.8**	95.0	**99.6**
Start Sn	71.1	62.0			69.3	61.4	52.8	39.4	46.7	37.8	9.9	4.0
Start Sp	69.9	**94.4**			70.9	**94.4**	78.2	**94.8**	77.6	**94.4**	54.2	**93.1**
Stop Sn	67.1	60.4			64.9	59.0	47.9	37.5	43.3	36.3	11.1	5.1
Stop Sp	88.6	**95.1**			89.6	**95.4**	94.4	**97.4**	94.4	**97.4**	94.1	**99.1**
D. melanogaster	All reported	High conf.	All reported	High conf.			All reported	High conf.	All reported	High conf.	All reported	High conf.
Intron Sn	79.8	74.6	72.8	62.6		*	66.2	54.3	49.7	34.4	35.8	20.9
Intron Sp	83.5	**98.9**	79.6	**98.8**			79.5	**98.8**	80.5	**99.0**	88.4	**99.5**
Start Sn	70.3	60.7	49.8	36.5			37.7	29.2	22.3	15.9	14.1	9.7
Start Sp	79.5	**97.4**	75.6	**96.7**			71.6	**95.6**	73.4	**94.5**	75.0	**93.5**
Stop Sn	75.3	68.4	56.7	45.2			44.7	36.9	26.7	19.8	15.8	11.2
Stop Sp	94.8	**99.3**	94.2	**98.8**			92.8	**98.5**	94.5	**98.9**	95.8	**99.2**
S. lycopersicum					All reported	High conf.			All reported	High conf.	All reported	High conf.
Intron Sn		*		*	80.6	76.8		*	76.0	69.4	46.4	25.7
Intron Sp					70.5	**92.0**			81.7	**93.5**	89.7	**95.6**
Start Sn					48.9	42.8			39.9	32.9	8.5	3.4
Start Sp					39.2	**72.4**			43.8	**74.6**	40.7	**77.9**
Stop Sn					51.9	46.6			42.3	35.6	10.1	4.9
Stop Sp					69.9	**83.6**			76.9	**85.5**	85.8	**92.0**
D. rerio					All reported	High conf.			All reported	High conf.	All reported	High conf.
Intron Sn		*		*	65.5	55.8		*	61.2	50.1	37.6	24.3
Intron Sp					84.4	**92.2**			86.8	**93.5**	90.1	**96.8**
Start Sn					47.6	39.6			39.6	31.3	14.3	8.7
Start Sp					61.4	**84.1**			70.4	**85.7**	64.1	**89.5**
Stop Sn					52.1	46.3			46.2	38.9	17.3	11.2
Stop Sp					79.8	**89.8**			85.6	**91.8**	87.6	**95.5**

* See the first column to the right

Table A.5: Accuracy assessment of GeneMark-ES, GeneMark-EP and GeneMark-EP+. GeneMark-EP+ was run with enforcement of (a) only high confidence intron hints, (b) only high confidence hints to gene starts and stops (c) enforcement of both (a) and (b). The accuracy is shown at gene level, exon level (for all exons and separately for the initial, internal, terminal, and single exons), intron level as well as for starts and stops. All the numbers were obtained for tests in genus-excluded mode.

		ES	EP	EP+ Introns	EP+ Starts / Stops	EP+ Full
Neurospora crassa	Gene Sn / Sp	64.5 / 71.9	64.7 / 72.0	66.0 / 73.6	66.3 / 73.4	67.1 / 74.5
	Exon Sn / Sp	**75.7 / 84.8**	**75.7 / 85.1**	**77.3 / 84.9**	**76.7 / 85.8**	**77.8 / 85.5**
	Initial Sn / Sp	70.9 / 81.3	70.2 / 81.2	72.4 / 82.0	72.2 / 82.3	73.2 / 82.7
	Internal Sn / Sp	77.4 / 88.2	77.2 / 89.1	80.4 / 87.5	77.8 / 89.8	80.4 / 88.7
	Terminal Sn / Sp	79.0 / 89.5	79.2 / 89.8	79.9 / 89.1	80.2 / 90.3	80.5 / 89.8
	Single Sn / Sp	74.7 / 70.7	74.5 / 69.9	73.3 / 71.5	75.7 / 70.5	74.3 / 71.7
	Intron Sn / Sp	79.5 / 89.8	79.6 / 90.5	82.5 / 90.4	80.2 / 90.9	82.5 / 90.7
	Start Sn / Sp	76.2 / 83.2	76.0 / 82.8	76.8 / 83.7	77.5 / 83.8	77.8 / 84.3
	Stop Sn / Sp	85.8 / 92.0	86.0 / 92.1	86.4 / 92.6	86.8 / 92.4	86.9 / 92.7
Caenorhabditis elegans	Gene Sn / Sp	46.8 / 46.4	45.2 / 42.8	46.4 / 45.0	46.3 / 43.8	47.4 / 45.8
	Exon Sn / Sp	**81.0 / 82.4**	**80.0 / 80.0**	**80.2 / 81.2**	**80.2 / 80.4**	**80.3 / 81.5**
	Initial Sn / Sp	53.5 / 63.4	53.1 / 60.1	53.3 / 61.7	53.8 / 60.8	54.0 / 62.4
	Internal Sn / Sp	90.7 / 87.7	89.6 / 86.4	89.9 / 87.1	89.6 / 86.7	89.8 / 87.4
	Terminal Sn / Sp	73.6 / 77.2	72.6 / 72.8	72.6 / 74.5	73.1 / 73.2	73.0 / 74.7
	Single Sn / Sp	15.6 / 50.5	16.6 / 46.5	16.7 / 48.3	18.1 / 47.4	17.8 / 48.8
	Intron Sn / Sp	87.5 / 86.4	86.4 / 84.4	86.7 / 85.5	86.4 / 84.7	86.6 / 85.7
	Start Sn / Sp	53.7 / 64.8	53.4 / 61.5	53.6 / 63.2	54.0 / 62.3	54.3 / 63.7
	Stop Sn / Sp	73.5 / 78.0	72.6 / 73.5	72.6 / 75.3	73.1 / 73.9	73.0 / 75.4
Arabidopsis thaliana	Gene Sn / Sp	55.8 / 54.0	57.5 / 55.4	65.2 / 63.1	65.7 / 61.4	73.2 / 69.1
	Exon Sn / Sp	**77.2 / 79.2**	**77.5 / 80.5**	**80.1 / 82.5**	**79.6 / 83.0**	**81.6 / 84.7**
	Initial Sn / Sp	60.5 / 68.9	61.1 / 69.5	63.3 / 71.4	66.7 / 73.9	67.9 / 75.3
	Internal Sn / Sp	87.1 / 83.4	87.3 / 85.1	90.6 / 87.1	87.6 / 87.3	90.5 / 89.2
	Terminal Sn / Sp	61.2 / 72.2	61.9 / 72.9	63.7 / 74.6	66.3 / 76.0	66.9 / 76.9
	Single Sn / Sp	58.6 / 74.2	59.1 / 73.3	58.3 / 76.1	64.3 / 75.9	63.7 / 78.5
	Intron Sn / Sp	85.2 / 82.4	85.5 / 83.9	89.0 / 86.3	86.0 / 85.5	89.0 / 87.7
	Start Sn / Sp	65.4 / 74.1	65.8 / 74.1	66.5 / 75.3	71.4 / 78.0	71.3 / 78.7
	Stop Sn / Sp	67.0 / 77.0	67.8 / 77.5	68.4 / 78.8	71.8 / 79.5	71.7 / 80.2
Drosophila melanogaster	Gene Sn / Sp	50.2 / 47.6	52.7 / 49.6	55.4 / 53.4	57.0 / 52.5	59.5 / 56.1
	Exon Sn / Sp	**67.6 / 72.0**	**68.3 / 74.6**	**70.9 / 76.6**	**69.8 / 76.3**	**71.9 / 78.1**
	Initial Sn / Sp	55.0 / 59.8	56.3 / 61.4	57.4 / 63.8	59.8 / 64.0	60.2 / 66.1
	Internal Sn / Sp	75.9 / 78.0	76.0 / 82.0	80.1 / 83.2	76.1 / 83.8	79.8 / 85.0
	Terminal Sn / Sp	63.1 / 68.2	64.4 / 69.6	65.5 / 72.4	67.2 / 71.5	67.7 / 73.7
	Single Sn / Sp	50.8 / 73.4	52.7 / 71.7	51.9 / 73.1	55.7 / 71.6	54.8 / 72.6
	Intron Sn / Sp	70.1 / 75.5	70.4 / 78.6	74.3 / 81.1	71.0 / 79.9	74.3 / 82.3
	Start Sn / Sp	58.4 / 65.5	59.7 / 66.6	60.1 / 68.5	63.3 / 68.9	63.1 / 70.4
	Stop Sn / Sp	68.5 / 75.3	69.5 / 75.9	70.0 / 78.2	72.1 / 76.9	72.0 / 78.7
Solanum lycopersicum	Gene Sn / Sp	19.3 / 16.1	24.1 / 19.5	31.3 / 25.7	28.9 / 22.4	36.3 / 28.9
	Exon Sn / Sp	**65.2 / 46.2**	**69.0 / 53.7**	**73.9 / 57.8**	**71.6 / 55.4**	**75.5 / 59.1**
	Initial Sn / Sp	40.2 / 31.1	44.1 / 33.7	47.0 / 36.4	50.7 / 36.9	51.8 / 38.8
	Internal Sn / Sp	79.0 / 51.4	82.3 / 62.9	88.5 / 67.5	82.7 / 64.6	88.4 / 69.1
	Terminal Sn / Sp	49.7 / 37.5	55.1 / 41.1	59.1 / 44.7	61.4 / 43.8	62.9 / 46.2
	Single Sn / Sp	29.7 / 46.9	33.2 / 42.1	32.7 / 43.6	37.1 / 43.9	36.6 / 45.2
	Intron Sn / Sp	71.9 / 48.8	76.3 / 58.7	84.6 / 65.1	77.2 / 59.7	84.6 / 65.9
	Start Sn / Sp	44.4 / 39.2	47.9 / 40.5	49.5 / 42.7	54.5 / 43.8	54.6 / 45.3
	Stop Sn / Sp	53.9 / 46.7	58.2 / 48.4	60.4 / 51.1	63.9 / 50.6	64.2 / 52.3
Danio rerio	Gene Sn / Sp	12.1 / 4.5	16.2 / 5.8	21.8 / 8.6	24.4 / 8.5	29.8 / 11.5
	Exon Sn / Sp	**64.0 / 43.7**	**66.5 / 48.0**	**71.2 / 52.4**	**68.8 / 50.3**	**72.7 / 54.2**
	Initial Sn / Sp	29.3 / 15.4	34.3 / 17.4	37.7 / 20.9	45.6 / 22.8	47.0 / 25.7
	Internal Sn / Sp	71.3 / 52.7	73.2 / 59.4	78.3 / 63.0	73.6 / 61.3	78.2 / 64.6
	Terminal Sn / Sp	43.8 / 23.2	47.9 / 24.5	51.4 / 28.8	55.5 / 28.1	56.5 / 31.2
	Single Sn / Sp	32.9 / 26.0	37.3 / 23.3	38.0 / 25.4	50.3 / 26.1	50.1 / 27.8
	Intron Sn / Sp	63.7 / 45.8	66.3 / 51.3	73.0 / 57.0	67.0 / 52.9	73.1 / 58.1
	Start Sn / Sp	32.9 / 17.5	38.3 / 19.5	40.9 / 22.8	51.3 / 25.5	51.7 / 28.1
	Stop Sn / Sp	48.7 / 26.0	52.6 / 26.9	55.6 / 31.0	61.7 / 30.8	62.0 / 33.8

Table A.6: Numbers of all annotated introns in the APPRIS set of principal isoforms and numbers of introns located within regions encoding conserved protein domains.

Species	Introns in the APPRIS set of principal isoforms		
	All	In regions coding for conserved domains	
D. melanogaster	41,010	21,562	(52.6%)
C. elegans	102,254	50,134	(49.0%)
D. rerio	178,867	106,288	(59.4%)

Table A.7: The change in the fraction of high-confidence and all reported intron hints mapped to conserved protein domains when the protein database size is changed from the largest (species or genus excluded) to the smallest (phylum excluded). Gene annotations use the principal protein isoforms defined by the APPRIS database.

Species	Exclusion level	High-confidence introns matching APPRIS introns			All reported introns matching APPRIS introns		
		All	In domains		All	In domains	
Drosophila melanogaster	Species	33,894	18,934	(55.9%)	35,338	19,414	(54.9%)
	Subgenus	28,437	17,475	(61.5%)	32,413	18,917	(58.4%)
	Family	24,670	16,057	(65.1%)	29,576	18,257	(61.7%)
	Order	15,829	11,984	(75.7%)	22,620	16,016	(70.8%)
	Phylum	9,719	8,222	(84.6%)	16,535	13,110	(79.3%)
Caenorhabditis elegans	Species	38,912	30,346	(78.0%)	80,402	45,210	(56.2%)
	Family	19,155	16,556	(86.4%)	39,379	29,270	(74.3%)
	Phylum	13,668	12,216	(89.4%)	27,464	23,140	(84.3%)
Danio rerio	Genus	108,236	71,239	(65.8%)	126,010	80,307	(63.7%)
	Order	97,457	67,335	(69.1%)	118,131	78,078	(66.1%)
	Phylum	47,860	40,117	(83.8%)	73,568	58,355	(79.3%)

APPENDIX B

BRAKER2

B.1 Extrinsic evidence configuration parameters in AUGUSTUS in BRAKER2

Extrinsic parameters for evidence integration with AUGUSTUS were adapted using *Arabidopsis thaliana* genome and hints generated with ProtHint using OrthoDB v10 Plants section (exempting proteins from the same species). Tests using other genomes and protein databases did not result in significantly different parameters. The final AUGUSTUS extrinsic parameters used for all species by BRAKER2 were:

```
[SOURCES]
M RM P C

# M: manual hints, to be enforced hints
# RM: repeats
# P: protein hints
# C: chained protein hints

[GENERAL]
start       1     1      M 1 1e+100 RM 1    1 P 2 1 1e3 1e6 C 1  1e6
stop        1     1      M 1 1e+100 RM 1    1 P 2 1 1e3 1e6 C 1  1e6
ass         1     1    1 M 1 1e+100 RM 1    1 P 2 1 1e2 1e2 C 1  1e2
dss         1     1    1 M 1 1e+100 RM 1    1 P 2 1 1e2 1e2 C 1  1e2
intron      1 0.168      M 1 1e+100 RM 1    1 P 2 1 1e2 100 C 1 3.16
CDSpart     1     1 0.99 M 1 1e+100 RM 1    1 P 2 1 1e2 1e4 C 1  1e4
nonexonpart 1     1      M 1 1e+100 RM 1 1.14 P 2 1   1   1 C 1    1
```

B.2 Running VARUS to sample and align RNA-Seq libraries

VARUS [94] (version from March 26, 2020) was run with fastq-dump [93] (v2.10.4) and HISAT2 [32] (v2.1.0). Results of VARUS depend on the date it was run because the amount of data deposited to NCBI Sequence Read Archive [93], from which VARUS samples the reads, is changing in time. Therefore, we uploaded the result of VARUS for each species at https://github.com/tomasbruna/braker2-exp/tree/master/${SPECIES}/varus. The aforementioned folder also contains information on when VARUS was run and what specific VARUS parameters (`VARUSparameters.txt`) were used.

B.3 Supplementary Figures

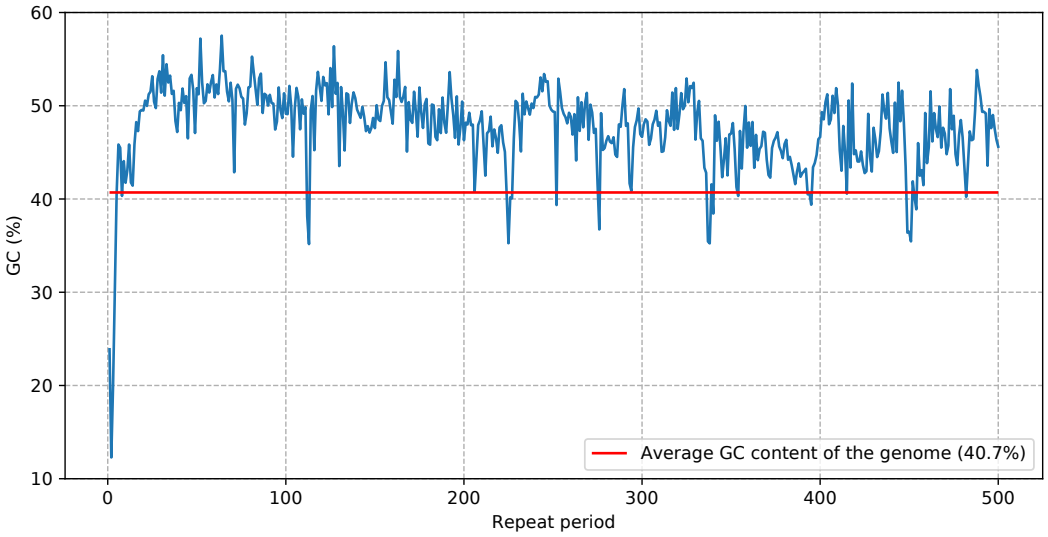

Figure B.1: GC-content of tandem repeats in the *X. tropicalis* genome shown as a function of the size of the repeat period.

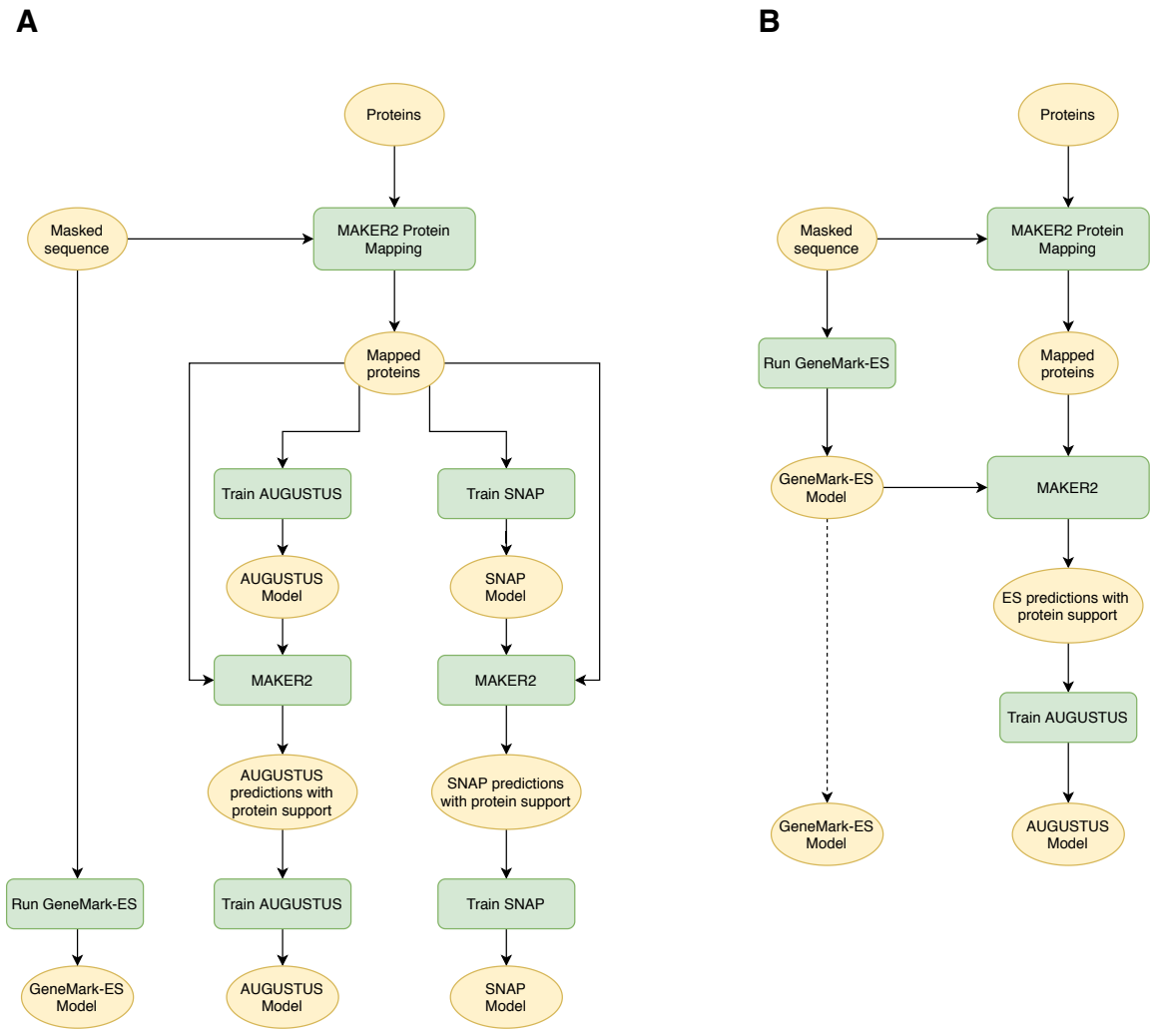

Figure B.2: Schematics of the MAKER2 training protocols: (A) a protocol recommended by the MAKER2 authors [121]; (B) an alternative protocol (conceptually similar to BRAKER2) that was implemented and produced better gene prediction accuracy.

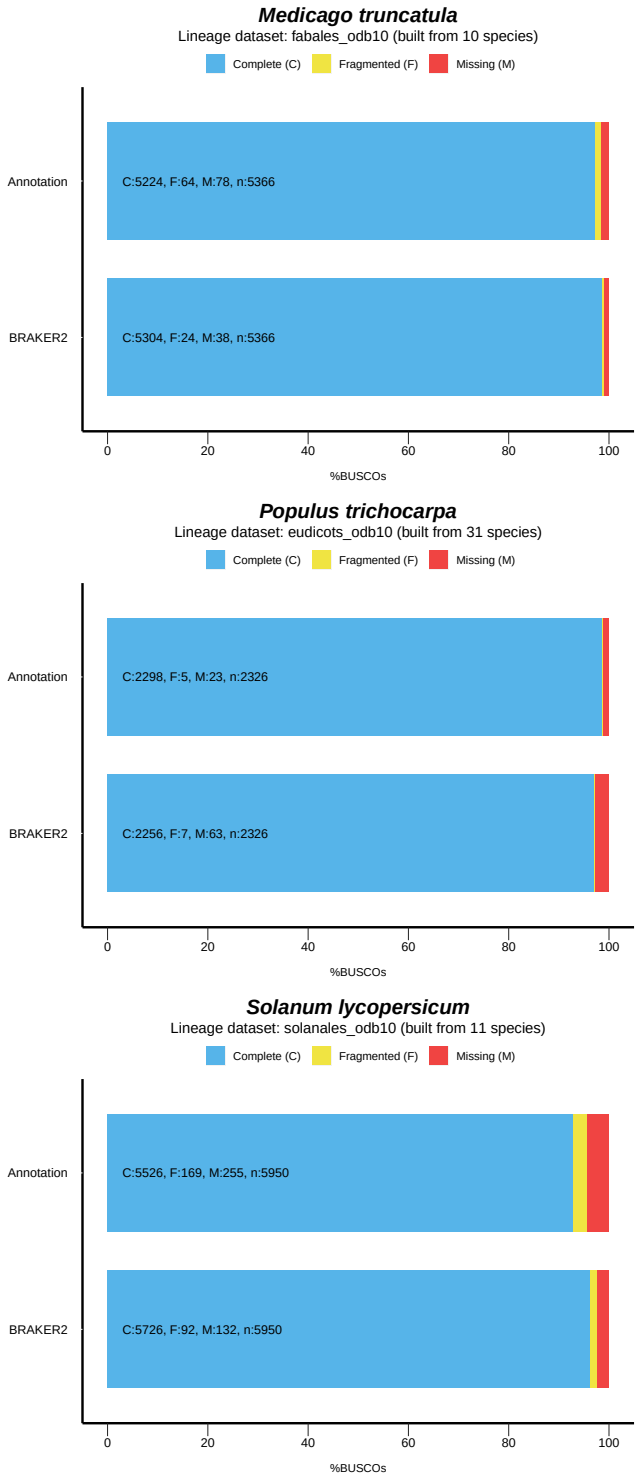

Figure B.3: Statistics of the sets of genes from BUSCO families (complete, fragmented, missing) of plant species identified in the reference genome annotation (top in each panel); the same statistics for the set of genes predicted by BRAKER2 (bottom in each panel).

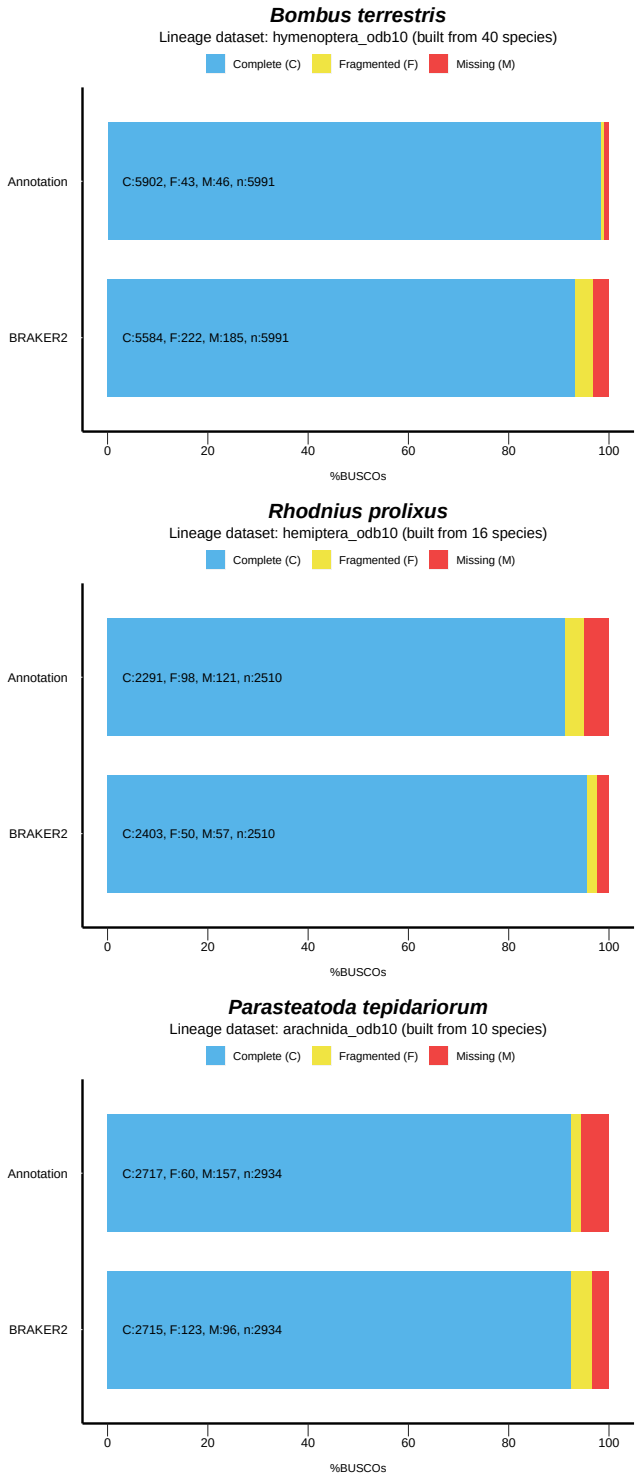

Figure B.4: BUSCO statistics for Arthropoda species. See the caption of Figure B.3 for details.

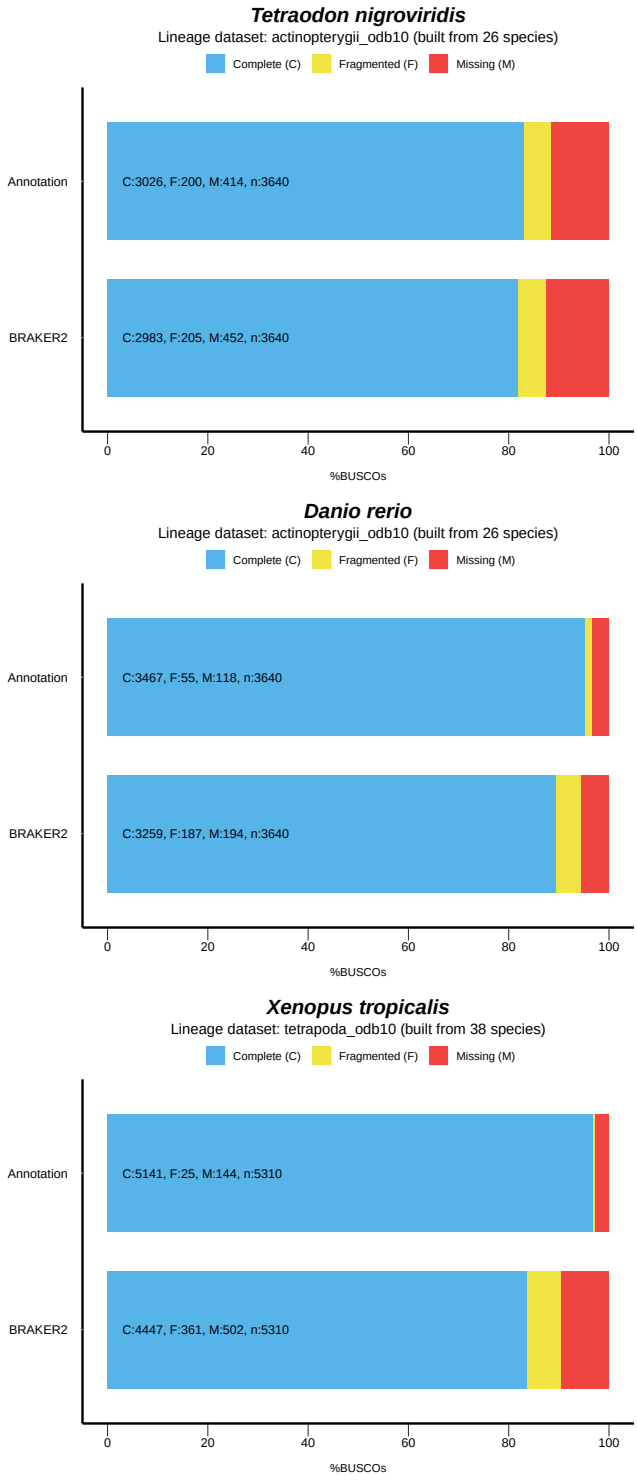

Figure B.5: BUSCO statistics for Metazoa species. See the caption of Figure B.3 for details.

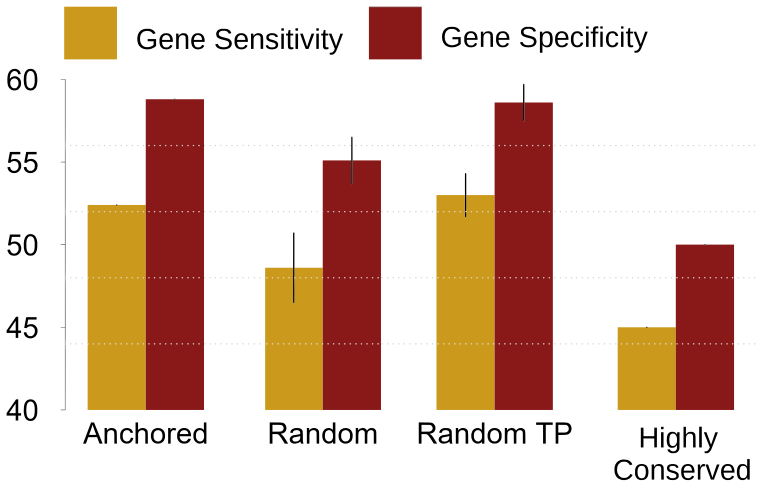

Figure B.6: The effect of selecting various AUGUSTUS training sets (selected from a GeneMark-EP+ prediction) on its *ab initio* prediction accuracy. See Section 4.3.5.4 for the description of this experiment. The genome of *D. melanogaster* with supporting proteins outside of the same phylum were used in this experiment.

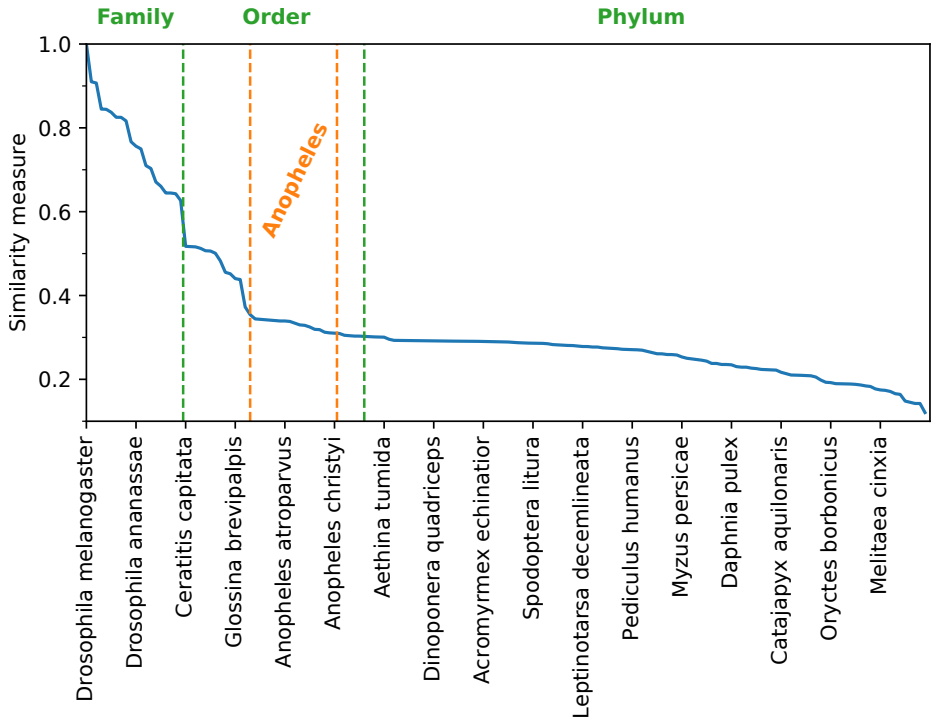

Figure B.7: Species selected for the accuracy evaluation experiment described in Section 4.3.5.5. The species are sorted in order of the increase of their evolutionary distance to *D. melanogaster* (the measure is defined in Section 4.3.5.5). In the X-axis, we show the name of every 10th species in the reference protein set. The green dashed lines separate species from inside and outside of the *D. melanogaster*'s taxonomic family (the left one), as well as the species from inside and outside of the *D. melanogaster*'s taxonomic order (the right one). The orange dashed lines delimit the space of the Anopheles species.

B.4 Supplementary Tables

Table B.1: Genome assemblies used for testing BRAKER2.

Species	Assembly version
Species with early sequenced genomes	
Arabidopsis thaliana	GCF_000001735
Caenorhabditis elegans	GCA_001483305
Drosophila melanogaster	GCA_000001215
Other species	
Plantae	
Populus trichocarpa	Ptrichocarpa_533_v4.0
Medicago truncatula	GCA_003473485.2
Solanum lycopersicum	SL4.0
Arthropoda	
Bombus terrestris	GCF_000214255.1
Rhodnius prolixus	GCA_000181055.3
Parasteatoda tepidariorum	GCF_000365465.2
Vertebrata	
Tetraodon nigroviridis	TETRAODON 8.0
Danio rerio	GCF_000002035
Xenopus tropicalis	GCF_000004195.4

Table B.2: Proteins of these species were used as external evidence in tests comparing MAKER2 with BRAKER2. The three groups of ten species were selected at random from the OrthoDB partitions (see Section 4.2.1).

A. thaliana	*C. elegans*	*D. melanogaster*
Dendrobium officinale	*Buceros rhinoceros silvestris*	*Pogonomyrmex barbatus*
Parasponia andersonii	*Cardiocondyla obscurior*	*Oryctes borbonicus*
Beta vulgaris subsp. vulgaris	*Drosophila elegans*	*Heliconius melpomene*
Aegilops tauschii	*Geospiza fortis*	*Stegodyphus mimosarum*
Nelumbo nucifera	*Sarcoptes scabiei*	*Calopteryx splendens*
Triticum urartu	*Austrofundulus limnaeus*	*Wasmannia auropunctata*
Ananas comosus	*Nomascus leucogenys*	*Fopius arisanus*
Coccomyxa subellipsoidea C-169	*Pieris rapae*	*Limulus polyphemus*
Populus euphratica	*Anas platyrhynchos*	*Tribolium castaneum*
Phalaenopsis equestris	*Numida meleagris*	*Myzus cerasi*

Table B.3: Gene prediction accuracy of BRAKER2 and BRAKER1 observed in tests on the *A. thaliana* genome. The sets of reference proteins for BRAKER2 were selected from the Plantae section of OrthoDB.

	BRAKER2			
	Order excluded	Family excluded	Species excluded	BRAKER1
Gene Sn	71.1	73.6	79.4	61.6
Gene Sp	67.0	69.7	72.8	61.7
Gene F1	69.0	71.6	76.0	61.6
Exon Sn	80.7	81.5	83.3	79.9
Exon Sp	86.6	87.4	86.8	81.7
Exon F1	83.5	84.3	85.0	80.8

Table B.4: The same information as in Table B.3 for a test on the *C. elegans* genome. The sets of reference proteins for BRAKER2 were selected from the Metazoa section of OrthoDB.

	BRAKER2			
	Order excluded	Family excluded	Species excluded	BRAKER1
Gene Sn	49.8	49.1	67.4	58.2
Gene Sp	56.2	55.1	68.3	62.3
Gene F1	52.8	51.9	67.8	60.2
Exon Sn	75.4	74.7	84.3	83.6
Exon Sp	88.6	88.2	90.7	87.2
Exon F1	81.5	80.9	87.4	85.4

Table B.5: The same information as in Table B.3 for a test on the *D. melanogaster* genome. The sets of reference proteins for BRAKER2 were selected from the Arthropoda section of OrthoDB.

	BRAKER2			
	Order excluded	Family excluded	Species excluded	BRAKER1
Gene Sn	61.1	66.3	77.8	63.1
Gene Sp	60.2	64.8	72.9	61.8
Gene F1	60.6	65.5	75.3	62.4
Exon Sn	71.4	74.5	79.8	76.7
Exon Sp	83.2	85.1	87.6	80.7
Exon F1	76.8	79.4	83.5	78.6

Table B.6: Complementary information for Table 4.3; Sn, Sp, and F1 values computed on exon and gene levels. Contrary to Table 4.3, the comparisons were made against the full complements of reference annotations; annotated single exons genes were included as well. For a gene to be considered complete and canonical, at least one of the gene's transcripts had to be fully annotated, such that the initial coding exon started with a "canonical" ATG and the terminal coding exon ended with TAA, TAG, or TGA.

Species	Gene			Exon			% Non-canonical or incomplete genes
	Sn	Sp	F1	Sn	Sp	F1	
P. trichocarpa	69.1	60.2	64.3	84.9	82.3	83.6	0.3
M. truncatula	44.7	44.0	44.3	78.7	71.5	74.9	0.0
S. lycopersicum	41.2	34.4	37.5	76.6	67.7	71.9	14.5
B. terrestris	46.9	25.0	32.6	74.5	72.0	73.2	4.7
R. prolixus	16.0	10.6	12.8	60.6	49.7	54.6	34.7
P. tepidariorum	30.4	14.9	20.0	67.7	59.6	63.4	18.2
T. nigroviridis	11.0	7.9	9.2	60.5	56.7	58.5	63.8
D. rerio	40.6	20.5	27.2	75.3	69.4	72.2	11.8
X. tropicalis	40.6	25.9	31.6	75.1	77.5	76.3	2.4

Table B.7: Numbers of genes, transcripts, and alternative transcripts predicted by BRAKER1 and BRAKER2 in genomes of three species with different sets of proteins on input (from the relevant OrthoDB partitions with proteins from the same species, family, and the order excluded).

		A. thaliana			
		Genes	Transcripts	Alternative transcripts	% Alt from all
	Annotation	27,444	40,827	13,383	32.8
	BRAKER1	27,403	28,899	1,496	5.2
BRAKER2 with exclusion of proteins from	Species	29,902	31,844	1,942	6.1
	Family	28,988	30,153	1,165	3.9
	Order	29,101	30,248	1,147	3.8

		C. elegans			
		Genes	Transcripts	Alternative transcripts	% Alt from all
	Annotation	20,172	28,506	8,334	29.2
	BRAKER1	18,833	20,978	2,145	10.2
BRAKER2 with exclusion of proteins from	Species	19,916	21,366	1,450	6.8
	Family	17,977	18,466	489	2.6
	Order	17,883	18,283	400	2.2

		D. melanogaster			
		Genes	Transcripts	Alternative transcripts	% Alt from all
	Annotation	13,929	22,247	8,318	37.4
	BRAKER1	14,208	15,470	1,262	8.2
BRAKER2 with exclusion of proteins from	Species	14,863	16,149	1,286	8.0
	Family	14,247	15,266	1,019	6.7
	Order	14,142	14,605	463	3.2

Table B.8: Accuracy of BRAKER2; determined for three genomes with different combinations of ProtHint hint types: high-confidence hints (HC), non-high-confidence hints (LC), chained CDSpart hints (Chains).

	C. elegans											
	Order excluded proteins				Family excluded proteins				Species excluded proteins			
	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains
Gene Sn	47.5	49.6	49.3	49.8	45.9	48.8	48.5	49.1	55.1	65.9	65.1	67.4
Gene Sp	54.9	56.6	55.9	56.2	52.8	55.2	54.6	55.1	58.7	67.1	66.4	68.3
Gene F1	50.9	52.9	52.4	52.8	49.1	51.8	51.4	51.9	56.9	66.5	65.7	67.8
Exon Sn	73.8	75.1	75.2	75.4	72.5	74.1	74.3	74.7	78.7	83.4	83.2	84.3
Exon Sp	88.7	89.0	88.7	88.6	88.0	88.5	88.3	88.2	89.4	91.0	90.9	90.7
Exon F1	80.6	81.5	81.4	81.5	79.5	80.7	80.7	80.9	83.7	87.0	86.9	87.4

	A.thaliana											
	Order excluded proteins				Family excluded proteins				Species excluded proteins			
	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains
Gene Sn	65.0	68.5	70.2	71.1	66.9	70.7	73.3	73.6	72.9	76.7	79.1	79.4
Gene Sp	63.6	66.0	66.4	67.0	65.8	68.5	69.6	69.7	69.6	71.6	73.2	72.9
Gene F1	64.3	67.2	68.3	69.0	66.3	69.6	71.4	71.6	71.2	74.1	76.0	76.0
Exon Sn	78.5	79.9	80.4	80.7	79.1	80.6	81.2	81.5	81.3	82.6	83.1	83.3
Exon Sp	86.3	86.6	86.6	86.6	87.0	87.4	87.6	87.4	86.8	86.6	87.2	86.7
Exon F1	82.2	83.1	83.4	83.6	82.9	83.8	84.3	84.3	83.9	84.5	85.1	85.0

	D. melanogaster											
	Order excluded proteins				Family excluded proteins				Species excluded proteins			
	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains	HC Hints	HC Hints LC Hints	HC Hints Chains	HC Hints LC Hints Chains
Gene Sn	58.6	60.2	60.4	61.1	62.8	64.6	65.7	66.3	73.7	76.0	77.5	77.8
Gene Sp	59.0	60.1	59.8	60.2	63.1	63.9	64.7	64.8	71.0	72.1	73.0	72.9
Gene F1	58.8	60.1	60.1	60.6	63.0	64.3	65.2	65.6	72.3	74.0	75.2	75.3
Exon Sn	69.4	70.5	70.9	71.4	72.2	73.5	74.0	74.5	78.0	79.1	79.5	79.8
Exon Sp	83.3	83.5	83.3	83.2	84.9	85.0	85.4	85.1	87.2	87.4	87.9	87.6
Exon F1	75.7	76.5	76.6	76.8	78.1	78.8	79.3	79.5	82.3	83.0	83.5	83.5

Table B.9: Change of the gene prediction accuracy upon successive steps of BRAKER2. Experiments on the three genomes used reference proteins from the relevant OrthoDB partitions with (A) proteins from the same taxonomic family excluded, and (B) proteins from the same species excluded.

A

A. thaliana	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Family excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	55.8	67.5	73.2	73.6
Gene Sp	54.0	64.6	69.4	69.7
Gene F1	54.9	66.0	71.3	71.6
Exon Sn	77.2	80.3	81.3	81.5
Exon Sp	79.2	83.7	87.3	87.4
Exon F1	78.2	81.9	84.2	84.3

C. elegans	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Family excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	46.8	47.4	48.9	49.1
Gene Sp	46.4	45.8	54.9	55.1
Gene F1	46.6	46.6	51.7	51.9
Exon Sn	81.0	80.3	74.7	74.7
Exon Sp	82.4	81.5	88.1	88.2
Exon F1	81.7	80.9	80.8	80.9

D. melanogaster	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Family excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	50.2	59.5	65.6	66.3
Gene Sp	47.6	56.1	64.1	64.8
Gene F1	48.9	57.7	64.8	65.6
Exon Sn	67.6	71.9	74.2	74.5
Exon Sp	72.0	78.2	84.9	85.1
Exon F1	69.7	74.9	79.2	79.5

B

A. thaliana	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Species excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	55.8	73.7	78.9	79.4
Gene Sp	54.0	69.4	72.7	72.9
Gene F1	54.9	71.5	75.7	76.0
Exon Sn	77.2	81.8	83.1	83.3
Exon Sp	79.2	84.8	86.8	86.7
Exon F1	78.2	83.2	84.9	85.0

C. elegans	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Species excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	46.8	53.4	66.8	67.4
Gene Sp	46.4	51.8	67.7	68.3
Gene F1	46.6	52.6	67.2	67.8
Exon Sn	81.0	82.4	84.1	84.3
Exon Sp	82.4	84.1	90.6	90.7
Exon F1	81.7	83.2	87.2	87.4

D. melanogaster	BRAKER2 steps 1 and 2		BRAKER2 steps 3 and 4	
Species excluded proteins	GeneMark		AUGUSTUS with hints	
	ES	EP+	first iteration	second iteration
Gene Sn	50.2	69.2	76.6	77.8
Gene Sp	47.6	63.1	72.0	72.9
Gene F1	48.9	66.0	74.2	75.3
Exon Sn	67.6	76.2	79.3	79.8
Exon Sp	72.0	80.9	87.3	87.6
Exon F1	69.7	78.5	83.1	83.5

Table B.10: Prediction accuracy of MAKER2 on three repeat-masked genomes. The table shows (i) the accuracy of gene finders trained directly on gene structures derived by protein alignments, as recommended by the MAKER2 protocol; (ii) the accuracy of gene finders trained on genes predicted by GeneMark-ES and supported at least partially by protein alignments (BRAKER2-like, see Figure B.2). Three combinations of gene finders in MAKER2 (SNAP + GeneMark-ES + AUGUSTUS; GeneMark-ES + AUGUSTUS; AUGUSTUS) are compared.

Species	*A. thaliana*			*C. elegans*			*D. melanogaster*		
Training	MAKER2 / BRAKER2-like			MAKER2 / BRAKER2-like			MAKER2 / BRAKER2-like		
Predictors	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS
Gene Sn	49.3 / 50.6	52.9 / **53.9**	48.5 / 49.8	25.5 / 26.2	28.4 / **30.4**	24.6 / 26.6	42.6 / 44.6	45.0 / **48.0**	42.8 / 46.2
Gene Sp	42.1 / 43.8	54.1 / **55.5**	49.9 / 51.8	22.1 / 23.0	37.1 / **38.9**	32.1 / 34.0	31.1 / 31.5	46.8 / **50.3**	44.8 / 48.8
Gene F1	45.4 / 47.0	53.5 / **54.7**	49.2 / 50.8	23.6 / 24.5	32.2 / **34.1**	27.9 / 29.8	35.9 / 37.0	45.9 / **49.2**	43.8 / 47.5
Exon Sn	73.4 / 73.8	74.5 / **74.7**	72.5 / 72.7	61.7 / **63.8**	59.7 / 62.6	58.3 / 61.2	62.8 / **64.3**	61.7 / 63.7	60.4 / 62.5
Exon Sp	72.6 / 72.9	**83.4** / 83.0	82.1 / 81.5	64.5 / 65.0	80.6 / **81.4**	78.3 / 79.2	58.7 / 54.6	75.3 / **76.0**	74.3 / 75.1
Exon F1	73.0 / 73.3	**78.7** / 78.6	77.0 / 76.8	63.1 / 64.4	68.6 / **70.8**	66.9 / 69.0	60.7 / 59.1	67.8 / **69.3**	66.6 / 68.2

Table B.11: The same comparison as in Table Table B.10, with gene predictions made on unmasked genomes.

Species	*A. thaliana*			*C. elegans*			*D. melanogaster*		
Training	MAKER2 / BRAKER2-like			MAKER2 / BRAKER2-like			MAKER2 / BRAKER2-like		
Predictors	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS	SNAP GM-ES AUGUSTUS	GM-ES AUGUSTUS	AUGUSTUS
Gene Sn	52.9 / 54.3	58.6 / **59.2**	53.5 / 54.7	34.9 / 34.5	43.0 / **43.9**	36.0 / 39.3	46.1 / 48.3	50.1 / **52.0**	45.8 / 49.3
Gene Sp	35.1 / 36.7	45.4 / 46.7	46.2 / **49.0**	25.8 / 25.5	38.1 / 38.9	43.7 / **46.8**	24.1 / 26.0	35.8 / 37.8	42.0 / **45.8**
Gene F1	42.2 / 43.8	51.2 / **52.2**	49.6 / 51.7	29.7 / 29.3	40.4 / 41.3	39.5 / **42.8**	31.7 / 33.8	41.8 / 43.8	43.8 / **47.5**
Exon Sn	75.7 / 76.0	**77.6** / 77.4	75.2 / 75.1	75.9 / 76.6	78.2 / **79.2**	69.2 / 72.5	65.7 / **67.1**	65.5 / 66.8	62.2 / 64.5
Exon Sp	62.8 / 63.3	72.3 / 72.1	**76.6** / 75.7	66.0 / 64.8	77.7 / 78.1	84.2 / **85.2**	46.5 / 46.2	60.0 / 60.8	69.8 / **69.7**
Exon F1	68.7 / 69.1	74.8 / 74.7	**75.9** / 75.4	70.6 / 70.2	77.9 / **78.7**	76.0 / 78.3	54.4 / 54.7	62.6 / 63.7	65.8 / **67.0**

APPENDIX C

GENEMARK-ETP+

C.1 Identification of GMS-T predictions that conflict with a cross-species protein alignment

GMS-T prediction candidates that are not fully supported by proteins must satisfy several conditions to be classified as high-confidence ones (Section 5.3.2.3). One such condition is that they must not conflict with a cross-species protein alignment. This condition is evaluated as follows: The prediction candidate is first mapped to the genomic DNA. Next, ProtHint (Section 3.3.2) is run using the mapped candidate as a gene seed. The elements of the mapped candidate (introns, start, and stop) not supported by ProtHint hints are selected for further inspection. Here, all the unsupported elements are compared to the spliced alignment of the best scoring homologous protein selected by ProtHint. If any of the unsupported elements conflict with this spliced alignment, the entire prediction candidate is deemed to conflict with a cross-species protein. An unsupported element is said to conflict with the spliced alignment when (i) an unsupported intron overlaps a spliced alignment-defined exon, (ii) an unsupported stop overlaps a spliced alignment-defined exon or intron, (iii) an unsupported start overlaps a spliced alignment-defined exon (except when the overlap coincides with the exon start) or intron.

C.2 Supplementary Figures

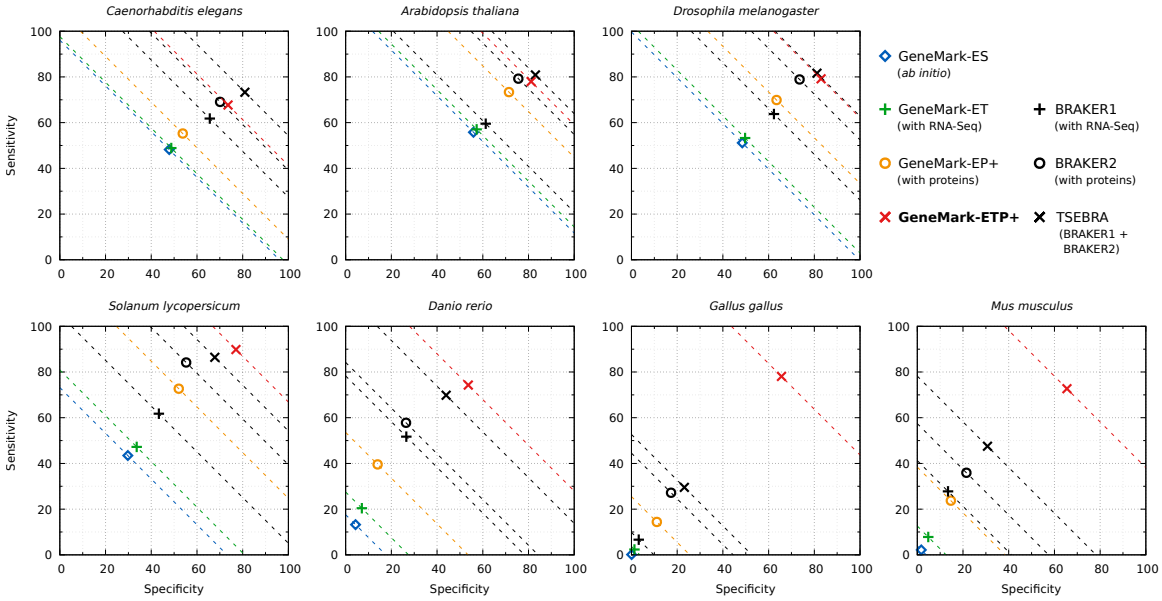

Figure C.1: Gene level accuracy of GeneMark-ETP+ and other tools. The comparisons are the same as in Figures 5.10 and 5.11; the only difference lies in the protein database used on input: all proteins except for the proteins belonging to the species of interest were used on input in all tests.

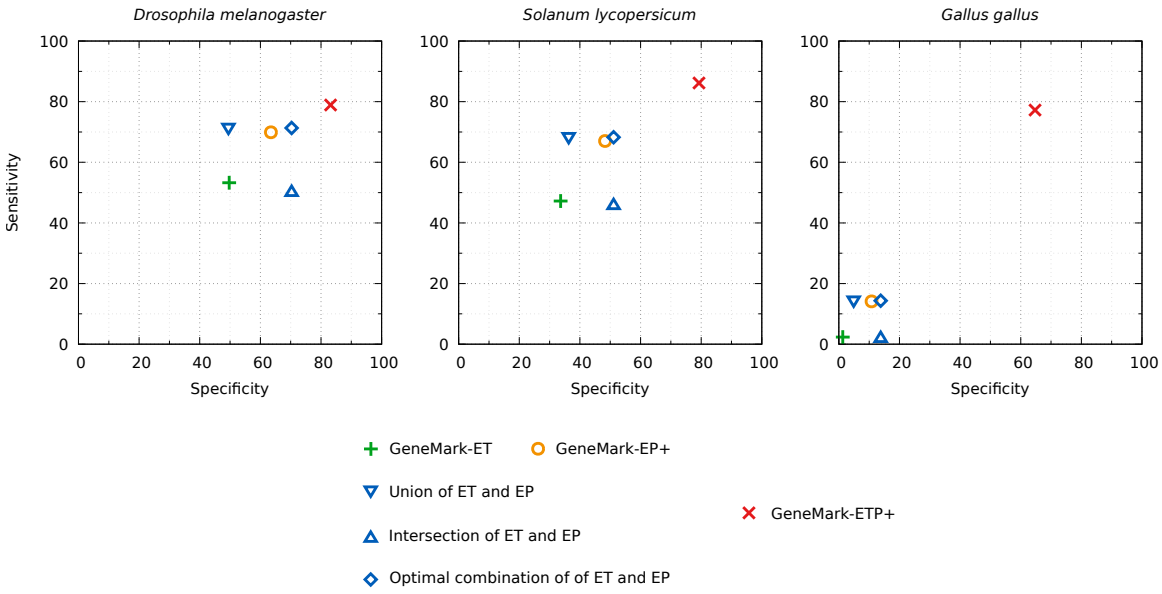

Figure C.2: Gene-level accuracy of the optimal combinations of GeneMark-ET (ET) and GeneMark-EP+ (EP+) (Section 5.3.5.3) compared to the prediction accuracy of GeneMark-ETP+. The results for *D. melanogaster* are shown with closely related proteins on input (only proteins of the tested species itself were excluded from the database); the other two genomes used remote proteins (proteins of the same taxonomic order as the species of interest removed from the input database).

C.3 Supplementary Tables

Table C.1: Data sources for each genome tested. The numbers in parentheses show the date of the last update.
(*) The reliable subset for *M. musculus* was selected by choosing a subset of GENCODE transcripts with the following attributes: `CCDS` (Agreement with RefSeq annotation), `transcript_support_level=1` (All splice junctions of the transcript are supported by at least one non-suspect mRNA), and `basic` (Prioritises full-length protein-coding transcripts over partial or non-protein-coding transcripts within the same gene).

Species	Assembly version	Main annotation	Supplementary annotation used to prepare the reliable subset
C. elegans	GCF_000002985.6	Wormbase WS284 (Feb 2022)	-
A. thaliana	GCF_000001735.4	Araport11 (Mar 2021)	-
D. melanogaster	GCF_000001215.4	FlyBase r6.44 (Feb 2022)	-
S. lycopersicum	GCF_000188115.4	NCBI annot. release 103 (Jun 2019)	ITAG3.2 (Jun 2017)
D. rerio	GCF_000002035.6	NCBI annot. release 106 (Oct 2019)	Ensembl GRCz11.105 (Oct 2021)
G. gallus	GCF_000002315.6	NCBI annot. release 104 (Mar 2020)	Ensembl GRCg6a.105 (Oct 2021)
M. musculus	GCF_000001635.27	GENCODE M28 (Dec 2021)	RefSeq*

Table C.2: Composition of the clades of OrthoDB v10.1 used by GeneMark-ETP+. Numbers in black bold show the largest numbers of species used to support gene predictions for a given species (left column). The numbers of species removed from the largest OrthoDB segment (see Section 5.2) are shown in blue.

Species	# of species in the OrthoDB clade						Name of the largest OrthoDB segment	# of proteins in the OrthoDB segment
	Genus	Family	Order	Class	Phylum	Kingdom		
C. elegans	**3**	3	**5**	6	7	**448**	Metazoa	8,266,016
A. thaliana	**2**	8	**10**	-	100	**117**	Plantae	3,510,742
D. melanogaster	**20**	20	**56**	148	**170**	-	Arthropoda	2,601,995
S. lycopersicum	**2**	10	**11**	-	100	**117**	Plantae	3,510,742
D. rerio	**1**	5	**5**	50	**246**	-	Chordata	5,003,104
G. gallus	**1**	3	**4**	62	246	-	Chordata	5,003,104
M. musculus	**3**	5	**20**	111	246	-	Chordata	5,003,104

Table C.3: RNA-Seq libraries used for the assessment of GeneMark-ETP+.

Species	RNA-Seq library ID	Number of paired reads (M)	Read length (nt)	Library size (Gb)
C. elegans	SRR065717	29.1	76	4.4
	SRR065719	73.3	76	11.1
	SRR473298	19.9	100	4.0
	SRR2054452	10.2	100	2.0
	Total	132.5		21.5
A.thaliana	SRR934391	20.0	101	4.0
	SRR5588566	24.7	125	6.2
	SRR7169927	19.2	101	3.9
	Total	63.9		14.1
D. melanogaster	SRR023505	8.4	76	1.3
	SRR023546	8.9	76	1.4
	SRR023608	11.9	76	1.8
	SRR026433	22.1	76	3.4
	SRR027108	7.2	76	1.1
	Total	58.5		9.0
S. lycopersicum	SRR2002284	56.2	73	8.2
	SRR7959012	25.4	149	7.6
	SRR7959019	27.9	149	8.3
	SRR14055940	21.2	150	6.4
	Total	130.7		30.5
D. rerio	SRR9735169	28.2	75	4.2
	SRR10004226	21.6	150	6.5
	SRR10040127	25.9	126	6.5
	Total	75.7		17.2
G. gallus	ERR2812450	44.9	150	13.5
	SRR3971633	24.0	100	4.8
	SRR6337028	10.0	100	2.0
	SRR11038071	16.4	151	5.0
	Total	95.3		25.3
M. musculus	SRR567480	155.7	101	31.5
	SRR567482	161.1	101	32.5
	SRR567497	94.3	101	19.0
	Total	411.1		83.0

Table C.4: Comparison of gene- and exon-level prediction accuracy between *ab initio* GeneMark-ES, RNA-Seq-based GeneMark-ET, protein-based GeneMark-EP, and GeneMark-ETP+. The accuracy estimates are shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

		ES	ET	Order-excluded		Species-excluded	
				EP+	ETP+	EP+	ETP+
C. elegans	Gene Sn	48.2	48.9	48.5	59.7	55.2	67.7
	Gene Sp	47.9	48.8	46.8	67.7	53.8	73.6
	Gene F1	48.0	48.8	47.6	63.5	54.5	70.5
	Exon Sn	81.8	81.7	81.1	82.0	83.3	85.1
	Exon Sp	83.1	83.7	82.0	90.4	84.9	91.5
	Exon F1	82.5	82.7	81.5	86.0	84.1	88.2
A. thaliana	Gene Sn	55.8	57.1	66.6	75.9	73.4	78.0
	Gene Sp	55.9	57.3	65.9	80.1	71.5	81.2
	Gene F1	55.9	57.2	66.3	77.9	72.4	79.5
	Exon Sn	76.9	77.1	79.8	82.2	81.5	82.9
	Exon Sp	80.8	82.1	84.9	91.0	86.3	91.1
	Exon F1	78.8	79.5	82.3	86.3	83.8	86.8
D. melanogaster	Gene Sn	51.2	53.3	56.5	71.6	69.9	79.2
	Gene Sp	48.5	49.7	53.9	78.2	63.5	83.0
	Gene F1	49.8	51.4	55.1	74.7	66.5	81.0
	Exon Sn	67.8	68.6	70.2	76.3	76.5	80.7
	Exon Sp	72.8	74.2	77.3	89.8	81.1	91.4
	Exon F1	70.2	71.3	73.6	82.5	78.8	85.7
S. lycopersicum	Gene Sn	43.4	47.2	67.0	87.4	72.7	89.8
	Gene Sp	29.8	33.6	48.3	79.2	52.0	77.1
	Gene F1	35.3	39.3	56.1	83.1	60.7	83.0
	Exon Sn	82.6	83.5	90.5	96.4	92.1	97.1
	Exon Sp	66.5	71.3	77.8	91.5	78.7	90.2
	Exon F1	73.7	76.9	83.7	93.9	84.9	93.5
D. rerio	Gene Sn	13.2	20.4	35.7	73.0	39.6	74.3
	Gene Sp	4.3	7.1	12.7	54.3	14.0	53.7
	Gene F1	6.5	10.5	18.7	62.3	20.7	62.3
	Exon Sn	75.3	79.1	84.9	93.6	86.2	94.0
	Exon Sp	39.6	48.9	54.3	83.4	54.9	82.9
	Exon F1	51.9	60.4	66.3	88.2	67.1	88.1
G. gallus	Gene Sn	0.1	2.4	14.1	78.4	14.4	78.1
	Gene Sp	0.1	1.3	10.8	66.7	11.1	65.6
	Gene F1	0.1	1.7	12.2	72.1	12.5	71.3
	Exon Sn	0.3	15.1	28.7	95.4	29.0	95.4
	Exon Sp	0.2	26.5	52.5	90.7	52.9	90.2
	Exon F1	0.2	19.2	37.1	93.0	37.5	92.8
M. musculus	Gene Sn	2.2	7.8	22.0	71.3	23.7	72.7
	Gene Sp	1.8	4.8	13.8	66.4	14.7	65.5
	Gene F1	1.9	6.0	17.0	68.7	18.1	68.9
	Exon Sn	25.4	49.7	57.3	90.8	58.1	91.5
	Exon Sp	25.1	50.3	63.7	91.8	64.3	91.5
	Exon F1	25.2	50.0	60.3	91.3	61.0	91.5

Table C.5: Comparison of gene- and exon-level prediction accuracy between RNA-Seq-based BRAKER1, protein-based BRAKER2, TSEBRA (a tool for the combination of BRAKER1 and BRAKER2 results), and GeneMark-ETP+. The accuracy estimates are shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

			Order-excluded			Species-excluded		
		BRAKER1	BRAKER2	TSEBRA	ETP+	BRAKER2	TSEBRA	ETP+
C. elegans	Gene Sn	61.8	46.8	62.8	59.7	69.1	73.3	67.7
	Gene Sp	65.6	54.1	78.2	67.7	70.1	81.0	73.6
	Gene F1	63.7	50.2	69.7	63.5	69.6	76.9	70.5
	Exon Sn	85.0	74.0	76.6	82.0	84.8	83.9	85.1
	Exon Sp	88.5	87.8	93.4	90.4	91.5	93.8	91.5
	Exon F1	86.7	80.3	84.2	86.0	88.0	88.6	88.2
A. thaliana	Gene Sn	59.6	72.6	74.8	75.9	79.3	80.9	78.0
	Gene Sp	61.3	70.1	81.4	80.1	75.6	83.0	81.2
	Gene F1	60.4	71.3	78.0	77.9	77.4	82.0	79.5
	Exon Sn	78.3	81.0	79.6	82.2	83.1	82.7	82.9
	Exon Sp	82.5	88.4	93.7	91.0	88.2	93.2	91.1
	Exon F1	80.4	84.5	86.1	86.3	85.6	87.6	86.8
D. melanogaster	Gene Sn	63.8	61.1	69.0	71.6	78.9	81.6	79.2
	Gene Sp	62.3	60.9	75.7	78.2	73.6	81.2	83.0
	Gene F1	63.0	61.0	72.2	74.7	76.1	81.4	81.0
	Exon Sn	77.0	71.4	72.1	76.3	80.1	79.8	80.7
	Exon Sp	80.9	83.4	89.9	89.8	88.5	92.2	91.4
	Exon F1	78.9	76.9	80.0	82.5	84.1	85.6	85.7
S. lycopersicum	Gene Sn	61.8	79.6	83.4	87.4	84.2	86.4	89.8
	Gene Sp	43.3	52.6	66.8	79.2	55.3	67.8	77.1
	Gene F1	50.9	63.3	74.2	83.1	66.7	76.0	83.0
	Exon Sn	90.7	94.2	94.9	96.4	95.4	96.1	97.1
	Exon Sp	73.1	80.6	88.5	91.5	80.4	88.3	90.2
	Exon F1	81.0	86.9	91.6	93.9	87.3	92.0	93.5
D. rerio	Gene Sn	51.7	55.0	67.7	73.0	57.8	69.8	74.3
	Gene Sp	26.6	28.0	43.8	54.3	26.4	44.0	53.7
	Gene F1	35.1	37.1	53.2	62.3	36.3	54.0	62.3
	Exon Sn	91.1	88.0	89.4	93.6	89.4	90.1	94.0
	Exon Sp	73.2	76.7	84.9	83.4	73.9	84.4	82.9
	Exon F1	81.2	81.9	87.1	88.2	80.9	87.1	88.1
G. gallus	Gene Sn	6.7	25.2	27.7	78.4	27.2	29.5	78.1
	Gene Sp	3.2	15.9	21.9	66.7	17.3	23.1	65.6
	Gene F1	4.3	19.5	24.4	72.1	21.2	25.9	71.3
	Exon Sn	66.1	35.0	59.8	95.4	35.3	60.0	95.4
	Exon Sp	47.3	58.3	73.3	90.7	59.6	73.2	90.2
	Exon F1	55.2	43.7	65.8	93.0	44.3	66.0	92.8
M. musculus	Gene Sn	27.8	32.5	44.9	71.3	35.9	47.5	72.7
	Gene Sp	13.5	19.6	29.4	66.4	21.5	30.7	65.5
	Gene F1	18.1	24.5	35.5	68.7	26.9	37.3	68.9
	Exon Sn	83.9	57.6	77.4	90.8	59.3	78.1	91.5
	Exon Sp	66.7	71.0	82.6	91.8	72.1	82.8	91.5
	Exon F1	74.3	63.6	79.9	91.3	65.1	80.4	91.5

Table C.6: A gene-level accuracy evaluation of raw GMS-T predictions and the final high-confidence (HC) genes. The accuracy is shown separately for complete and incomplete predictions as well as for both sets together (Combined). The first three columns (Raw GMS-T) show the prediction accuracy of unprocessed GMS-T predictions in all assembled transcripts. The remaining columns (HC genes) show the accuracy of the processed, high-confidence gene sets. The accuracy of HC genes is shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

		Raw GMS-T			HC genes					
					Order-excluded proteins			Species-excluded proteins		
		Complete	Incomplete	Combined	Complete	Incomplete	Combined	Complete	Incomplete	Combined
C. elegans	Sn	43.7	3.9	47.6	33.6	2.1	35.8	48.0	4.0	52.0
	Sp	82.2	18.3	63.8	88.9	81.5	88.4	91.6	80.7	90.6
A. thaliana	Sn	50.3	1.4	51.7	55.9	1.1	57.0	57.6	1.6	59.1
	Sp	89.2	17.0	80.0	97.4	92.3	97.3	97.8	90.8	97.6
D. melanogaster	Sn	57.2	3.2	60.5	53.4	1.8	55.2	61.0	3.1	64.1
	Sp	87.7	38.1	82.0	95.0	85.3	94.7	96.5	84.8	95.9
S. lycopersicum	Sn	66.6	1.4	68.0	73.9	1.3	75.1	74.4	1.5	75.8
	Sp	81.2	21.0	74.6	93.1	81.3	92.8	92.8	78.9	92.4
D. rerio	Sn	56.2	4.3	60.5	63.2	4.2	67.3	63.4	4.5	68.0
	Sp	66.0	29.0	57.2	86.0	73.8	84.6	86.0	70.0	84.0
G. gallus	Sn	44.1	5.7	49.8	68.2	6.5	74.7	66.6	7.7	74.3
	Sp	61.0	18.3	43.3	86.0	82.8	85.6	86.6	75.7	84.8
M. musculus	Sn	48.8	1.2	50.0	61.0	2.7	63.7	60.8	3.4	64.2
	Sp	76.1	7.5	59.5	92.4	62.7	90.4	92.4	65.3	90.3

Table C.7: Accuracy of the complete/incomplete classification (described in Section 5.3.2.2). The transcripts shown in this evaluation were classified as incomplete by GMS-T, had a correctly predicted stop codon, and contained no assembly errors. The row and column names are the same as in the confusion matrix shown in Figure 5.13, see Section 5.4.3 for details. Sensitivity represents the percentage of complete transcripts that were classified as such. Error rate represents the percentage of incomplete transcripts that were incorrectly classified as complete. The results are shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

		Order-excluded proteins		Species-excluded proteins	
		Actually complete	Truly incomplete	Actually complete	Truly incomplete
C. elegans	Predicted complete	1487	127	1981	78
	Predicted incomplete	274	207	394	471
	Sensitivity	84.4		83.4	
	Error rate	38.0		14.2	
A. thaliana	Predicted complete	1476	55	1442	22
	Predicted incomplete	107	203	165	249
	Sensitivity	93.2		89.7	
	Error rate	21.3		8.1	
D. melanogaster	Predicted complete	272	77	299	9
	Predicted incomplete	49	253	130	388
	Sensitivity	84.7		69.7	
	Error rate	23.3		2.3	
S. lycopersicum	Predicted complete	920	90	887	69
	Predicted incomplete	80	314	118	349
	Sensitivity	92.0		88.3	
	Error rate	22.3		16.5	
D. rerio	Predicted complete	1159	122	1054	80
	Predicted incomplete	210	1262	325	1345
	Sensitivity	84.7		76.4	
	Error rate	8.8		5.6	
G. gallus	Predicted complete	3318	227	3034	128
	Predicted incomplete	456	892	741	999
	Sensitivity	87.9		80.4	
	Error rate	20.3		11.4	
M. musculus	Predicted complete	1945	9	1799	3
	Predicted incomplete	472	209	620	218
	Sensitivity	80.5		74.4	
	Error rate	4.1		1.4	

Table C.8: Comparison of gene- and exon-level prediction accuracy between GeneMark-ETP+ results with and without filtering of pure *ab initio* prediction. The superior F1 accuracy between the two sets is highlighted in bold. The pure *ab initio* predictions are removed from the default GeneMark-ETP+ output in genomes $\geq$ 300 Mbp in size (the bottom four genomes). The results are shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

		Order-excluded		Species-excluded	
		All predictions	Pure *ab initio* removed	All predictions	Pure *ab initio* removed
C. elegans	Gene Sn	59.7	58.1	67.7	67.0
	Gene Sp	67.7	71.0	73.6	75.8
	Gene F1	63.5	**63.9**	70.5	**71.1**
	Exon Sn	82.0	80.2	85.1	84.6
	Exon Sp	90.4	91.8	91.5	92.4
	Exon F1	**86.0**	85.6	88.2	**88.3**
A. thaliana	Gene Sn	75.9	71.7	78.0	77.5
	Gene Sp	80.1	86.4	81.2	84.1
	Gene F1	77.9	**78.4**	79.5	**80.7**
	Exon Sn	82.2	80.8	82.9	82.6
	Exon Sp	91.0	94.5	91.1	92.6
	Exon F1	86.3	**87.1**	86.8	**87.3**
D. melanogaster	Gene Sn	71.6	64.2	79.2	78.6
	Gene Sp	78.2	82.0	83.0	84.8
	Gene F1	**74.7**	72.0	81.0	**81.6**
	Exon Sn	76.3	74.0	80.7	80.5
	Exon Sp	89.8	92.7	91.4	93.0
	Exon F1	**82.5**	82.3	85.7	**86.3**
S. lycopersicum	Gene Sn	89.1	87.4	90.3	89.8
	Gene Sp	68.3	79.2	68.7	77.1
	Gene F1	77.4	**83.1**	78.1	**83.0**
	Exon Sn	96.8	96.4	97.2	97.1
	Exon Sp	85.9	91.5	85.9	90.2
	Exon F1	91.0	**93.9**	91.2	**93.5**
D. rerio	Gene Sn	73.3	73.0	74.4	74.3
	Gene Sp	38.9	54.3	39.0	53.7
	Gene F1	50.8	**62.3**	51.2	**62.3**
	Exon Sn	93.8	93.6	94.2	94.0
	Exon Sp	73.3	83.4	73.2	82.9
	Exon F1	82.3	**88.2**	82.3	**88.1**
G. gallus	Gene Sn	78.6	78.4	78.2	78.1
	Gene Sp	42.8	66.7	42.5	65.6
	Gene F1	55.4	**72.1**	55.1	**71.3**
	Exon Sn	95.5	95.4	95.4	95.4
	Exon Sp	79.2	90.7	79.0	90.2
	Exon F1	86.6	**93.0**	86.5	**92.8**
M. musculus	Gene Sn	71.7	71.3	72.8	72.7
	Gene Sp	35.3	66.4	35.9	65.5
	Gene F1	47.3	**68.7**	48.1	**68.9**
	Exon Sn	91.2	90.8	91.7	91.5
	Exon Sp	72.3	91.8	72.6	91.5
	Exon F1	80.7	**91.3**	81.0	**91.5**

Table C.9: The masking penalty values estimated by GeneMark-ETP+ for each of the tested species. In GC-heterogeneous genomes, GeneMark-ETP+ estimates an optimal masking penalty for each of the GC bins. The values are shown in the logarithmic space (natural logarithm). The results are shown for two different protein sets used on input: remotely related (proteins of the same taxonomic order as the species of interest removed from the input database) and closely related (only the proteins of the species of interest removed).

	Order-excluded proteins			Species-excluded proteins		
C. elegans	0.06			0.05		
A. thaliana	0.03			0.03		
D. melanogaster	0.09			0.08		
S. lycopersicum	0.04			0.04		
D. rerio	0.08			0.07		
	GC			GC		
	Low	Medium	High	Low	Medium	High
G. gallus	0.15	0.17	0.12	0.14	0.17	0.12
M. musculus	0.13	0.14	0.14	0.13	0.14	0.14

Table C.10: Percentages of sequence masked by RepeatModeler2/RepeatMasker in each tested genome.

Species	% Genome masked
C. elegans	18.7
A. thaliana	17.2
D. melanogaster	19.3
S. lycopersicum	62.4
D. rerio	60.4
G. gallus	13.1
M. musculus	42.0

REFERENCES

[1] Victoria Dominguez Del Angel et al. “Ten steps to get started in Genome Assembly and Annotation”. In: *F1000Research* 7 (2018).

[2] Katharina J Hoff and Mario Stanke. “Current methods for automated annotation of protein-coding genes”. In: *Current Opinion in insect science* 7 (2015), pp. 8–14.

[3] Mark Yandell and Daniel Ence. “A beginner’s guide to eukaryotic genome annotation”. In: *Nature Reviews Genetics* 13.5 (2012), pp. 329–342.

[4] Harris A. Lewin et al. “Earth BioGenome Project: Sequencing life for the future of life”. In: *Proceedings of the National Academy of Sciences* 115.17 (2018), pp. 4325–4333.

[5] Eric W Sayers et al. “GenBank”. In: *Nucleic Acids Research* 48.D1 (Oct. 2019), pp. D84–D86.

[6] NCBI. *GenBank’s Genome Table*. https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/. Accessed: 2022-03-05. 2022.

[7] Tatiana Tatusova et al. “NCBI prokaryotic genome annotation pipeline”. In: *Nucleic acids research* 44.14 (2016), pp. 6614–6624.

[8] Steven L Salzberg. “Next-generation genome annotation: we still struggle to get it right”. In: *Genome biology* 20.1 (2019), pp. 1–3.

[9] Hyungtaek Jung et al. “Twelve quick steps for genome assembly and annotation in the classroom”. In: *PLoS Computational Biology* 16.11 (2020), e1008325.

[10] David Kulp David Haussler and Martin G Reese Frank H Eeckman. “A generalized hidden Markov model for the recognition of human genes in DNA”. In: *Proc. int. conf. on intelligent systems for molecular biology, st. louis*. 1996, pp. 134–142.

[11] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[12] Mario Stanke et al. “AUGUSTUS: a web server for gene finding in eukaryotes”. In: *Nucleic acids research* 32.suppl_2 (2004), W309–W312.

[13] Ian Korf. “Gene finding in novel genomes”. In: *BMC bioinformatics* 5.1 (2004), pp. 1–9.

[14] Chris Burge and Samuel Karlin. “Prediction of complete gene structures in human genomic DNA”. In: *Journal of molecular biology* 268.1 (1997), pp. 78–94.

[15] Victor Solovyev et al. “Automatic annotation of eukaryotic genes, pseudogenes and promoters”. In: *Genome biology* 7.1 (2006), pp. 1–12.

[16] Alexandre Lomsadze et al. “Gene identification in novel eukaryotic genomes by self-training algorithm”. In: *Nucleic acids research* 33.20 (2005), pp. 6494–6506.

[17] Vardges Ter-Hovhannisyan et al. “Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training”. In: *Genome research* 18.12 (2008), pp. 1979–1990.

[18] Fawaz Ghali et al. “ProteoAnnotator–open source proteogenomics annotation software supporting PSI standards”. In: *Proteomics* 14.23-24 (2014), pp. 2731–2741.

[19] Nicolas Scalzitti et al. “A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms”. In: *BMC genomics* 21.1 (2020), pp. 1–20.

[20] Stephen J Goodswen, Paul J Kennedy, and John T Ellis. “Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques”. In: *PLoS One* 7.11 (2012), e50609.

[21] Avril Coghlan et al. “nGASP–the nematode genome annotation assessment project”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–13.

[22] Roderic Guigo et al. “EGASP: the human ENCODE genome annotation assessment project”. In: *Genome biology* 7.1 (2006), pp. 1–31.

[23] Michael Lynch. “The origins of eukaryotic gene structure”. In: *Molecular biology and evolution* 23.2 (2006), pp. 450–468.

[24] Lee P Lim and Christopher B Burge. “A computational analysis of sequence features involved in recognition of short introns”. In: *Proceedings of the National Academy of Sciences* 98.20 (2001), pp. 11193–11198.

[25] Hiroshi Akashi. “Gene expression and molecular evolution”. In: *Current opinion in genetics & development* 11.6 (2001), pp. 660–666.

[26] Samuel Karlin and Jan Mrázek. “Compositional differences within and between eukaryotic genomes”. In: *Proceedings of the National Academy of Sciences* 94.19 (1997), pp. 10227–10232.

[27] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[28] Rory Stark, Marta Grzelak, and James Hadfield. “RNA sequencing: the teenage years”. In: *Nature Reviews Genetics* 20.11 (2019), pp. 631–656.

[29] Brian J Haas et al. “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”. In: *Nature protocols* 8.8 (2013), pp. 1494–1512.

[30] Marcel H Schulz et al. “Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels”. In: *Bioinformatics* 28.8 (2012), pp. 1086–1092.

[31] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[32] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature biotechnology* 37.8 (2019), pp. 907–915.

[33] Mihaela Pertea et al. “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”. In: *Nature biotechnology* 33.3 (2015), pp. 290–295.

[34] Sam Kovaka et al. “Transcriptome assembly from long-read RNA-seq alignments with StringTie2”. In: *Genome biology* 20.1 (2019), pp. 1–13.

[35] Li Song et al. “A multi-sample approach increases the accuracy of transcript assembly”. In: *Nature communications* 10.1 (2019), pp. 1–7.

[36] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[37] Miten Jain et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome biology* 17.1 (2016), pp. 1–11.

[38] John Eid et al. “Real-time DNA sequencing from single polymerase molecules”. In: *Science* 323.5910 (2009), pp. 133–138.

[39] Thomas D Wu and Colin K Watanabe. “GMAP: a genomic mapping and alignment program for mRNA and EST sequences”. In: *Bioinformatics* 21.9 (2005), pp. 1859–1875.

[40] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (2018), pp. 3094–3100.

[41] Jason L Weirather et al. “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. In: *F1000Research* 6 (2017).

[42] Rachael E Workman et al. “Nanopore native RNA sequencing of a human poly (A) transcriptome”. In: *Nature methods* 16.12 (2019), pp. 1297–1305.

[43] Shiyuyun Tang, Alexandre Lomsadze, and Mark Borodovsky. “Identification of protein coding regions in RNA transcripts”. In: *Nucleic acids research* 43.12 (2015), e78–e78.

[44] Tamara Steijger et al. “Assessment of transcript reconstruction methods for RNA-seq”. In: *Nature methods* 10.12 (2013), pp. 1177–1184.

[45] Francisco Pardo-Palacios et al. “Systematic assessment of long-read RNA-seq methods for transcript identification and quantification”. In: (2021).

[46] Krešimir Križanović et al. “Evaluation of tools for long read RNA-seq splice-aware alignment”. In: *Bioinformatics* 34.5 (2018), pp. 748–754.

[47] Katharina J Hoff et al. “BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS”. In: *Bioinformatics* 32.5 (2016), pp. 767–769.

[48] Alexandre Lomsadze, Paul D Burns, and Mark Borodovsky. “Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm”. In: *Nucleic acids research* 42.15 (2014), e119–e119.

[49] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[50] Mikhail S Gelfand, Andrey A Mironov, and Pavel A Pevzner. “Gene recognition via spliced sequence alignment”. In: *Proceedings of the National Academy of Sciences* 93.17 (1996), pp. 9061–9066.

[51] Guy St C Slater and Ewan Birney. “Automated generation of heuristics for biological sequence comparison”. In: *BMC bioinformatics* 6.1 (2005), pp. 1–11.

[52] Gordon Gremme et al. “Engineering a software tool for gene structure prediction in higher organisms”. In: *Information and Software Technology* 47.15 (2005), pp. 965–978.

[53] Boris Kiryutin, Alexandre Souvorov, and Tatiana Tatusova. “ProSplign–Protein to Genomic Alignment Tool”. In: *Proc. 11th Annual International Conference in Research in Computational Molecular Biology, San Francisco, USA*. 2007.

[54] Ewan Birney, Michele Clamp, and Richard Durbin. “GeneWise and genomewise”. In: *Genome research* 14.5 (2004), pp. 988–995.

[55] Oliver Keller et al. “A novel hybrid gene prediction method employing protein multiple sequence alignments”. In: *Bioinformatics* 27.6 (2011), pp. 757–763.

[56] Osamu Gotoh. “Direct mapping and alignment of protein sequences onto genomic sequence”. In: *Bioinformatics* 24.21 (2008), pp. 2438–2444.

[57] Jens Keilwagen et al. “Using intron position conservation for homology-based gene prediction”. In: *Nucleic acids research* 44.9 (2016), e89–e89.

[58] Jens Keilwagen et al. “Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi”. In: *BMC bioinformatics* 19.1 (2018), pp. 1–12.

[59] Samuel S Gross et al. “CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction”. In: *Genome biology* 8.12 (2007), pp. 1–16.

[60] Stefanie Konig et al. “Simultaneous gene finding in multiple genomes”. In: *Bioinformatics* 32.22 (2016), pp. 3388–3395.

[61] Catherine Mathé et al. “Current methods of gene prediction, their strengths and weaknesses”. In: *Nucleic acids research* 30.19 (2002), pp. 4103–4117.

[62] Moises Burset and Roderic Guigo. “Evaluation of gene structure prediction programs”. In: *genomics* 34.3 (1996), pp. 353–367.

[63] Stefanie Konig, Lars Romoth, and Mario Stanke. “Comparative genome annotation”. In: *Comparative Genomics* (2018), pp. 189–212.

[64] Rong She et al. “genBlastG: using BLAST searches to build homologous gene models”. In: *Bioinformatics* 27.15 (2011), pp. 2141–2143.

[65] Bronwen L Aken et al. “The Ensembl gene annotation system”. In: *Database* 2016 (2016).

[66] Evgenia V Kriventseva et al. “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs”. In: *Nucleic acids research* 47.D1 (2019), pp. D807–D811.

[67] Evgeny M Zdobnov et al. “OrthoDB in 2020: evolutionary and functional annotations of orthologs”. In: *Nucleic acids research* 49.D1 (2021), pp. D389–D393.

[68] Jaime Huerta-Cepas et al. “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses”. In: *Nucleic acids research* 47.D1 (2019), pp. D309–D314.

[69] “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D480–D489.

[70] Lars Gabriel et al. “TSEBRA: transcript selector for BRAKER”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–12.

[71] Sagnik Banerjee et al. “FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–26.

[72] David E Cook et al. “Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing”. In: *Plant physiology* 179.1 (2019), pp. 38–54.

[73] Jinhwa Kong et al. “GAAP: a genome assembly+ annotation pipeline”. In: *BioMed research international* 2019 (2019).

[74] Franziska Zickmann and Bernhard Y Renard. “IPred-integrating ab initio and evidence based gene predictions to improve prediction accuracy”. In: *BMC genomics* 16.1 (2015), pp. 1–8.

[75] Qian Liu et al. “Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction”. In: *Bioinformatics* 24.5 (2008), pp. 597–605.

[76] Brian J Haas et al. “Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments”. In: *Genome biology* 9.1 (2008), pp. 1–22.

[77] Jonathan E Allen and Steven L Salzberg. “JIGSAW: integration of multiple sources of evidence for gene prediction”. In: *Bioinformatics* 21.18 (2005), pp. 3596–3603.

[78] Jonathan E Allen, Mihaela Pertea, and Steven L Salzberg. “Computational gene prediction using multiple sources of evidence”. In: *Genome Research* 14.1 (2004), pp. 142–148.

[79] Kevin L Howe, Tom Chothia, and Richard Durbin. “GAZE: a generic framework for the integration of gene-prediction data by dynamic programming”. In: *Genome research* 12.9 (2002), pp. 1418–1427.

[80] Sylvain Foissac et al. “Genome annotation in plants and fungi: EuGene as a model platform”. In: *Current Bioinformatics* 3.2 (2008), pp. 87–97.

[81] Erika Sallet, Jérôme Gouzy, and Thomas Schiex. “EuGene: an automated integrative gene finder for eukaryotes and prokaryotes”. In: *Gene Prediction*. Springer, 2019, pp. 97–120.

[82] Jonas Behr et al. “Next generation genome annotation with mGene. ngs”. In: *BMC bioinformatics* 11.10 (2010), pp. 1–2.

[83] Osamu Gotoh, Mariko Morita, and David R Nelson. “Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment”. In: *Bmc Bioinformatics* 15.1 (2014), pp. 1–13.

[84] Genis Parra, Keith Bradnam, and Ian Korf. “CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes”. In: *Bioinformatics* 23.9 (2007), pp. 1061–1067.

[85] Jose Manuel Rodriguez et al. “APPRIS 2017: principal isoforms for multiple gene sets”. In: *Nucleic acids research* 46.D1 (2018), pp. D213–D217.

[86] Jose Manuel Rodriguez et al. “APPRIS: annotation of principal and alternative splice isoforms”. In: *Nucleic acids research* 41.D1 (2013), pp. D110–D117.

[87] Benjamin Buchfink, Chao Xie, and Daniel H Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature methods* 12.1 (2015), pp. 59–60.

[88] Steven Henikoff and Jorja G Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919.

[89] Alexander V Lukashin and Mark Borodovsky. “GeneMark. hmm: new solutions for gene finding”. In: *Nucleic acids research* 26.4 (1998), pp. 1107–1115.

[90] Mark Borodovsky and James McIninch. “GENMARK: parallel gene recognition for both DNA strands”. In: *Computers & chemistry* 17.2 (1993), pp. 123–133.

[91] Arian Smit and Robert Hubley. *RepeatModeler Open-1.0*. http://www.repeatmasker.org/. 2008–2015.

[92] Arian Smit, Robert Hubley, and P. Green. *RepeatMasker Open-4.0*. http://www.repeatmasker.org/. 2013–2015.

[93] Rasko Leinonen et al. “The sequence read archive”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D19–D21.

[94] Mario Stanke et al. “VARUS: sampling complementary RNA reads from the sequence read archive”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–7.

[95] Christiam Camacho et al. “BLAST+: architecture and applications”. In: *BMC bioinformatics* 10.1 (2009), pp. 1–9.

[96] Aron Marchler-Bauer et al. “CDD/SPARCLE: functional classification of proteins via subfamily domain architectures”. In: *Nucleic acids research* 45.D1 (2017), pp. D200–D203.

[97] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[98] Michael L Tress, Federico Abascal, and Alfonso Valencia. “Most alternative isoforms are not functionally important”. In: *Trends in biochemical sciences* 42.6 (2017), pp. 408–410.

[99] Michael L Tress, Federico Abascal, and Alfonso Valencia. “Alternative splicing may not be the key to proteome complexity”. In: *Trends in biochemical sciences* 42.2 (2017), pp. 98–110.

[100] Vladimir Shulaev et al. “The genome of woodland strawberry (Fragaria vesca)”. In: *Nature genetics* 43.2 (2011), pp. 109–116.

[101] Shuai Zhan et al. “The monarch butterfly genome yields insights into long-distance migration”. In: *Cell* 147.5 (2011), pp. 1171–1185.

[102] Huajun Zheng et al. “The genome of the hydatid tapeworm Echinococcus granulosus”. In: *Nature genetics* 45.10 (2013), pp. 1168–1175.

[103] Huangwei Chu et al. “The floral organ number4 gene encoding a putative ortholog of Arabidopsis CLAVATA3 regulates apical meristem size in rice”. In: *Plant physiology* 142.3 (2006), pp. 1039–1052.

[104] Rafal Woycicki et al. “The genome sequence of the North-European cucumber (Cucumis sativus L.) unravels evolutionary adaptation mechanisms in plants”. In: *PloS one* 6.7 (2011), e22728.

[105] Mario Stanke et al. “AUGUSTUS: ab initio prediction of alternative transcripts”. In: *Nucleic acids research* 34.suppl_2 (2006), W435–W439.

[106] Mario Stanke et al. “Using native and syntenically mapped cDNA alignments to improve de novo gene finding”. In: *Bioinformatics* 24.5 (2008), pp. 637–644.

[107] Mario Stanke et al. “Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources”. In: *BMC bioinformatics* 7.1 (2006), pp. 1–11.

[108] Katharina J Hoff and Mario Stanke. "WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes". In: *Nucleic acids research* 41.W1 (2013), W123–W128.

[109] Derek M Bickhart et al. "Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome". In: *Nature genetics* 49.4 (2017), pp. 643–650.

[110] Yuki Yoshida et al. "Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus". In: *PLoS biology* 15.7 (2017), e2002266.

[111] John L Bowman et al. "Insights into land plant evolution garnered from the Marchantia polymorpha genome". In: *Cell* 171.2 (2017), pp. 287–304.

[112] José F Munoz et al. "Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species". In: *Nature communications* 9.1 (2018), pp. 1–13.

[113] Charissa De Bekker et al. "Ant-infecting Ophiocordyceps genomes reveal a high diversity of potential behavioral manipulation genes and a possible major role for enterotoxins". In: *Scientific reports* 7.1 (2017), pp. 1–13.

[114] Carson Holt and Mark Yandell. "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects". In: *BMC bioinformatics* 12.1 (2011), pp. 1–14.

[115] Mathieu Seppey, Mose Manni, and Evgeny M Zdobnov. "BUSCO: assessing genome assembly and annotation completeness". In: *Gene prediction*. Springer, 2019, pp. 227–245.

[116] Katharina J Hoff and Mario Stanke. "Predicting genes in single genomes with AUGUSTUS". In: *Current protocols in bioinformatics* 65.1 (2019), e57.

[117] Gary Benson. "Tandem repeats finder: a program to analyze DNA sequences". In: *Nucleic acids research* 27.2 (1999), pp. 573–580.

[118] Felipe A Simao et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19 (2015), pp. 3210–3212.

[119] Sean R Eddy. "Accelerated profile HMM searches". In: *PLoS computational biology* 7.10 (2011), e1002195.

[120] Brandi L Cantarel et al. “MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes”. In: *Genome research* 18.1 (2008), pp. 188–196.

[121] Michael S Campbell et al. “Genome annotation and curation using MAKER and MAKER-P”. In: *Current protocols in bioinformatics* 48.1 (2014), pp. 4–11.

[122] Oliver Keller et al. “Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–12.

[123] Genis Parra, Enrique Blanco, and Roderic Guigo. “GeneID in drosophila”. In: *Genome research* 10.4 (2000), pp. 511–515.

[124] Ole K Tørresen et al. “Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases”. In: *Nucleic acids research* 47.21 (2019), pp. 10994–11006.

[125] Philipp E Bayer, David Edwards, and Jacqueline Batley. “Bias in resistance gene prediction due to repeat masking”. In: *Nature Plants* 4.10 (2018), pp. 762–765.

[126] Shu Ouyang and C Robin Buell. “The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants”. In: *Nucleic acids research* 32.suppl_1 (2004), pp. D360–D363.

[127] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. “Optimization by simulated annealing”. In: *science* 220.4598 (1983), pp. 671–680.

[128] Jullien M Flynn et al. “RepeatModeler2 for automated genomic discovery of transposable element families”. In: *Proceedings of the National Academy of Sciences* 117.17 (2020), pp. 9451–9457.

[129] Arian Smit, Robert Hubley, and P. Green. *RepeatMasker 4.1.0*. http://www.repeatmasker.org/. 2013–2019.

[130] Wei Zhao et al. “Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling”. In: *BMC genomics* 15.1 (2014), pp. 1–11.

[131] Marilyn Kozak. “Initiation of translation in prokaryotes and eukaryotes”. In: *Gene* 234.2 (1999), pp. 187–208.

[132] Marilyn Kozak. “An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs”. In: *Nucleic acids research* 15.20 (1987), pp. 8125–8148.