GeneMark-EP and -EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins

Tomas Bruna[1,†], Alexandre Lomsadze[2†,] and Mark Borodovsky[1,2,3,*]

[1]School of Biological Sciences, [2]Wallace H. Coulter Department of Biomedical Engineering, [3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta GA 30332, USA

* To whom correspondence should be addressed: borodovsky@gatech.edu
† Joint first authors

Abstract
We have made several steps towards creating fast and accurate algorithm for gene prediction in eukaryotic genomes. First, we introduced an automated method for efficient *ab initio* gene finding, GeneMark-ES, with parameters trained in iterative *unsupervised* mode. Next, in GeneMark-ET we proposed a method of integration of unsupervised training with information on intron positions revealed by mapping short RNA reads. Now we describe GeneMark-EP, a tool that utilizes another source of external information, a protein database, readily available prior to a start of a sequencing project. The new algorithm and software tool integrate information produced by proteins spliced aligned to genomic regions into model training and gene prediction steps. A specialized pipeline, ProtHint, makes processing the results of mapping of multiple proteins to a genomic region where a protein from the same family is likely encoded. GeneMark-EP uses the hints from ProtHint to improve estimation of model parameters as well as to adjust co-ordinates of predicted genes if they disagree with the most reliable hints (the -EP+ mode). Tests conducted with GeneMark-EP and -EP+ have demonstrated that the gene prediction accuracy is higher than one of GeneMark-ES, particularly in large eukaryotic genomes.

Introduction

One of major challenges of gene prediction in eukaryotes is finding an optimal way to combine extrinsic and intrinsic sources of information. External information could be transferred from RNA transcripts as well as from cross-species proteins either derived from annotated genomes or determined by proteomics. Integration of transcript information, e.g. RNA-Seq reads, with ab initio gene prediction was implemented in several algorithms and software tools, e.g. AUGUSTUS (1), GeneMark-ET (2), EuGene (3,4), mGene.ngs (5). Also, a few other tools made use of protein sequences. The task of transferring protein information for gene identification in a newly sequenced genome is complex. Therefore, mapping a single protein to genomic locus where a homologous protein could be encoded was considered to be a separate task and specialized tools were developed for protein spliced alignment (e.g. currently available GeneWise (6), GenomeThreader (7), ProSplign (8), Spaln (9)). Notably, whole families of homologous proteins could be used to map elements of gene and protein structure conserved in evolution, e.g. AUGUSTUS-PPX (10) that used protein profiles derived for conserved protein domains. Information about conservation of intron position with respect to protein primary structures of multiple homologs was used in another tool, GeMoMa (11). Notably, an attempt to combine protein profiles with intron position conservation for assessment and refinement of predicted eukaryotic genes was made upon construction of yet another method GSA-MPSA (12).

A weakness of methods heavily relying on mapping of homologous proteins is the patchiness of this evidence; a sizable fraction of the whole complement of genes may code for proteins with few or no orthologues. Another weakness is that protein splice alignments become less accurate as the distance between the two species increases. Therefore, ab initio gene finders (e.g. GENSCAN (13), GeneMark.hmm (14), AUGUSTUS (15) or GeneID (16)) have been a necessary part of genome annotation tools and pipelines (e.g. GNOMON (17), PASA (18) and Ensembl (19)).

Application of *ab initio* algorithms for genome wide eukaryotic gene prediction was for long time hampered by the need of tedious and time-consuming training. To address this issue we have earlier developed an *ab initio* gene finder GeneMark-ES (20,21) with model parameters estimated by iterative unsupervised training. This algorithm

did not require expert based training or hints for building a training set. GeneMark-ET (2) was an extension of GeneMark-ES developed for integration of transcript information, raw RNA-Seq reads spliced aligned to genome in question.

Here we describe GeneMark-EP, an algorithm and software tool using external information extracted from reference set of cross-species protein sequences. To generate protein hints for a given genomic locus we first identify a set of proteins, homologous to the true protein encoded in a genomic locus. Then a specialized pipeline, ProtHint, computes the hints, a set of mapped splice sites (intron borders) and translation start and stop sites with scores characterizing hint confidence. The most reliably constructed elements of spliced alignment could be used to directly identify elements of exon-intron structures, this mode of algorithm execution with direct gene structure correction we call GeneMark-EP+.

A key question is how to find optimal method of hint incorporation into the *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 weights to protein hints may generate parameters dictated by a narrow set of conserved genes and proteins (22). By design, the GeneMark-EP algorithm combines strong features of both methods: i/ ability of unsupervised iterative training of an *ab initio* gene finder to create a set of training sequences with a size beyond reach of conventional supervised training and ii/ ability to correct model parameters and structures of newly discovered genes with respect to splice alignments of homologous cross-species proteins. The new method falls into category of gene prediction methods with semi-supervised training.

Materials

For assessment of GeneMark-EP as well as ProtHint accuracy we selected annotated genomes from diverse clades - fungi, worms, plants, insects, and vertebrae (Table 1). The genome length varied from under 100 Mb (*Neurospora crassa*) to more than 1.3 Gb (*Danio rerio*). With exception of *Solanum lycopersicum*, a species representating economically important plants with long genomes, all selected species are model organisms whose genomes presumably have high quality annotation. Therefore, to assess accuracy of gene prediction for such species, we compared genes predicted and annotated on a whole genome scale. In case of *S. lycopersicum* we used a limited set of genes, validated by available RNA-Seq data. In all genomic datasets, isolated contigs were excluded from the analysis as well as genomes of organelles.

We used OrthoDB v10 protein database (23) as an all-inclusive source of protein sequences. However, for generating protein hints for particular species we used subsets of OrthoDB: plant proteins for *Arabidopsis thaliana*, arthropod proteins for *Drosophila melanogaster*, etc. (Table 2).

As an additional test set we used annotation of major protein isoforms available in the APPRIS database (24); this assessment was done for *C. elegans*, *D. melanogaster*, *and D. rerio* (Table S1). Accuracy of prediction of major isoforms may be of special interest. Notably, in each gene locus a major isoform is expressed in higher volume than other (minor isoforms) (24).

Methods

Integration of genomic sequence patterns and protein homology into gene prediction

The GeneMark-EP, -EP+ algorithm goes step-by-step through the following tasks: i/ selection of genomic regions, *seed regions*, containing gene candidates (*the seed genes)*; ii/ identification for each seed region a set of homologous proteins iii/ processing splice alignments of homologous proteins to each seed region and creating hints for exon-intron structure; iv/ running itereative semi-supervised training with selection of most reliable elements of predicted genes in each iteration; v/ gene prediction with an option (-EP+ mode) of enforcing high confidence hints in predicted exon-intron structures (Fig. 1).

The first three tasks i/-iii/ are devoted to generating protein hints and are resolved by the ProtHint pipeline (Fig. 2). Particularly, to determine *seed regions* within a long genomic sequence (task i/) we run unsupervised training of GeneMark-ES models (20) and generate *ab initio* gene predictions. Each predicted gene, the *seed gene,* is expanded upstream and downstream by 2,000 nt margins to create a *seed region*. To identify proteins homologous to a *seed protein*, task ii/, we run DIAMOND similarity search (25) with a *seed protein* as a query against a protein sequence database (e.g. a section of OrthoDB). A set of proteins with statistically significant hits define a set of

target proteins presumed to be homologous to the query, the seed protein. The task iii/ is to generate spliced alignments of multiple protein targets to the seed region (done by either Spaln (9) or ProSplign (8)) and to process the alignments results to infer elements of exon-intron structures (introns, splice sites, translation starts and stops) characterized by reliability scores. Mapped gene elements with reliability scores exceeding chosen thresholds are designated as high-confidence hints. The final tasks (iv) and (v) correspond to training and prediction steps of GeneMark-EP and -EP+. At these steps we use the extrinsic hints to exon-intron structure co-ordinates as an input to an expectation-maximization type algorithm that simultaneously finds compositional patterns of protein-coding and non-coding regions along with the most likely parse of genomic sequence into coding and non-coding regions.

Iterative training of the GeneMark-EP statistical model (tasks iv and v) works as follows. In the first iteration splice sites and introns mapped by ProtHint with scores exceeding a stringent threshold (high confidence elements) are used to estimate parameters of splice sites models as well as branch point site models (particularly important for intron models of fungal genomes). The site models together with the heuristic models of protein-coding and non-coding regions make a complete set of models of a semi-Markov HMM (20). The models are used in the first run of the Viterbi-like algorithm (see (14)) that generates a parse of genomic sequence into coding and non-coding regions, the first set of genes predicted by GeneMark-EP. Next, we analyze available data to make updated training sets and re-estimate model parameters. We compare co-ordinates of exons predicted by GeneMark.hmm and exons determined by ProtHint within the *seed* regions. This comparison leads to selection of 'anchored' elements, the exons with at least one splice site identified by both GeneMark.hmm and ProtHint. A set of anchored exons along with a set of predicted single exon genes (with length > 800nt) comprise an updated training set for the three-matrix model of protein-coding region (26). Sequences of introns bounded by two anchored splice sites as well as intergenic sequences bordered by anchored terminal and initial exons of adjacent genes (Fig.3) are used for updating parameters of the non-coding region model. The set of 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 uniformed ones. More accurate estimation of these distributions is done in subsequent steps of iterative training (Fig. 1). Also, in the later steps we estimate parameters of the three-phase models of splice sites indexed by a nucleotide position after which the intron divides a codon triplet. In the final iterations we update estimates of the HMM transition probabilities that affect frequencies of genes with specific number of introns. Experimental runs done for genomes of different length were made to verify that the seven iterations are sufficient for GeneMark-ES and six iterations for GeneMark-EP and -EP+ to reach convergence in terms of co-ordinates of predicted genes and values of model parameters. Gene predictions made with the final model are reported output of GeneMark-EP.

Running the Viterbi algorithm (in logarithmic mode) could be done with enforcing high confidence elements mapped by ProtHint. Particularly, it is done by modifying components of the object functions of the Viterbi algorithm associated with chosen hidden states (sites). The sites that must be enforced receive high values of 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 execution of GeneMark-EP we call GeneMark-EP+.

ProtHint: generating footprints (hints) of multiple homologous proteins for a genomic locus

***General logic*.** The ProtHint role (Fig. 2) in GeneMark-EP, - EP+ is two-fold. This pipeline generates two sets of protein hints. The smaller one, the set of high confidence hints, includes hints with high scores that ensure their high specificity. The larger one includes hints that have scores exceeding a liberally set threshold, thus these hints have lower specificity but larger sensitivity. In the process of hint generation ProtHint takes a *seed protein* and uses it as a query in similarity search for homologs of a true protein presumably encoded in the seed region. Next, ProtHint constructs spliced alignments of the detected homologs (target proteins) to the seed region. The whole set of multiple spliced alignments is then processed together to identify the protein hints, mapped co-ordinates of the candidate splice sites, translation start and stop sites. Hints scoring system is discussed in detail in Supplementary Materials.

Technically, for a given *seed pro*tein, ProtHint runs DIAMOND (25) against a relevant section of the OrthoDB database and retains in the output up to 25 target proteins (with hit E-value better than 0.001). Next, the target

proteins are spliced aligned by Spaln (9) back to the seed region. Notably, the hints are defined by ProtHint processing Spaln raw pairwise alignments rather than by annotation of exons in the Spaln output. Multiple target proteins spliced aligned to a given seed region may map out the same sequence fragment as an intron. Such an outcome defines an intron hint with a higher confidence than if an intron candidate is mapped only once.

Score system for introns. As described above the expected evolutionary conservation between the structures of target proteins and the protein encoded in the seed region has to be quantified and used for accurate identification of the new gene. To facilitate this quantification, we define three types of scores for introns and adjacent exons (AEE, IBA and IMC, see below) and two types of scores for initial and terminal exons (SMC and BAQ, see below).

Alignment of Entire Exon (AEE) score is defined as a score of the Spaln (or ProSplign) alignment of exon translation and a target protein (see Supplementary Materials).

Intron Borders Alignment (IBA) score is computed for two adjacent exons with more weight given to parts close to the splice sites (within a window of length *w*). The score is computed as follows.

For downstream (and upstream) exon defined in the Spaln spliced alignment we compute S_d (and S_u) as

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

Here $S_a(G_i, P_i)$ is a substitution score for target protein amino acid P_i and a codon defined amino acid G_i; $W(i)$ is the weight function. For instance, for a downstream exon S_d:

$$W_i = \frac{K(i)}{\sum_{i=1}^{w} K(i)} \qquad (2)$$

where $K(i)$ is the kernel value for position i counting in codons from a splice site. In a linear kernel:

$$K(i) = 1 - \frac{|i| - 1}{w} \qquad (3)$$

Then we take a geometric mean of values of S_d and S_u.

$$S_{intron} = \begin{cases} \sqrt{S_u \times S_d}, & \min(S_u, S_d) > 0 \\ 0, & otherwise \end{cases} \qquad (4)$$

Finally, a joint IBA score is obtained by normalizing the S_{intron} score into $\langle 0, 1 \rangle$ range:

IBA score $= S_{intron} \,/\, \max(S_a)$, where $\max(S_a)$ is a maximum score among elements of the BLOSUM62 matrix

Intron Mapping Coverage (IMC) score counts how many times an intron was exactly mapped in spliced alignments of target proteins. Notably, introns with identical coordinates are represented by a single intron characterized by the maximum of individual IBA scores among all collapsed introns.

Application of the intron scores. For target proteins mapped into a particular seed region we use the three types of scores (AEE, IBA and IMC) as follows:

a/ we select introns with both upstream and downstream exons having AEE $\geq E_t$ selected as a threshold. For E_t =25 we observed relatively high Sn value of the candidate introns (Fig. S1). Further increase of E_t eliminated true introns while not significantly improving Sp value**.**

b/ from the set of selected introns we further choose ones with IBA score > I_t. For I_t = 0.1 we observed further increase Sp of the candidate introns without noticeable change in Sn (Fig. S1).

Thus identified set of introns represents a set of ***all*** *mapped introns*; it is used as external evidence to generate *anchored introns* for GeneMark-EP training steps as described above.

Within the set of ***all*** *mapped introns* we select a narrower set of *high-confidence introns*. These introns must have canonical GT-AG splice sites, an IMC score ≥ 4, and an IBA score ≥ 0.25 (Figs. 4, S2). Notably, the IMC score is computed only from the set of ***all*** *mapped introns*.

We use high-confidence introns to estimate initial parameters of the GeneMark-EP intron model. Importantly, these introns are enforced in the prediction steps of all iterations in GeneMark-EP+ mode.

Score system for translation starts and stops. Similarly, to scores introduced for intron mapping we define a *Border Alignment Quality* (BAQ) score for initial and terminal exons. This score is computed for *w* amino acids downstream (upstream) of start (stop) codon, weighted by a linear function (Eq. 1). The second type of score is the *Site Mapping Coverage* (SMC) score. This score is a count of N-terminals (C-terminals) of target proteins aligned to a particular start (stop) codon position of a candidate gene.

If a set of target proteins for a given seed region generates footprints situated more upstream than others, alternative start candidates situated downstream are removed from consideration (Fig. S3, details in Supplementary Materials). We have observed that using these rules leads to increase in accuracy (Table 3, S2).

Application of start and stop scores. All over, selection of a set of ***all*** translation start and stop hints is done by the following rules:

a/ stop codon candidates are selected from alignments of terminal exons of target proteins to stop codons in the seed region; start codon candidates are selected as ATG in an initial exon that aligns to N-terminal methionine in a target protein.

b/ the candidate initial (terminal) exons should have AEE score $E_t \geq 25$ and BAQ score ≥ 0.

To select a narrower set of *high-confidence hints* we choose stop codon hints with SMC score ≥ 4 as well as start codon hints with SMC score ≥ 4 *and* no overlap by longer target proteins. The SMC scores are computed only from the set of ***all*** *translation starts and stops*. This set of *high-confidence hints* is used to estimate parameters of GeneMark-EP models of translation initiation and termination sites. Also, the *high-confidence hints* are directly enforced in prediction steps GeneMark-EP+.

Do introns mapped by ProtHint tend to hit gene regions coding for conserved domains?

To address this question we use the following procedure. Annotated genes are translated to proteins and used as queries in RPS-BLAST (27) to search (E-value =0.01) against NCBI Conserved Domains Database (CDD) (28). Results of the RPS-BLAST searches are processed with *rpsbproc* utility (28) to generate a map of conserved domains for each query. Finally, coordinates of the conserved domains are mapped back to a *seed region* of genomic DNA and compare with ProtHint output to find out how many introns are 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 (24) as genes coding for principal protein isoforms (see Results).

Assessment of GeneMark-EP gene merging and gene splitting errors

Instances of gene splitting could be due to the current GeneMark-ES, -EP, -EP+ algorithms design for predicting non-overlapping genes with no alternative isoforms. Some types of errors are unavoidable due to this algorithm setting. Therefore, we excluded some genes from the test set: i/ genes fully overlapping shorter genes present inside introns situated in the same or in the opposite strand; ii/ genes with isoforms combining shorter alternative components (Fig. S4); iii/ genes with introns longer than 10,000nt (the default maximum intron length). For genes with annotated multiple alternative isoforms we used the longest one as a representative. Overlapping genes present in annotation (e.g. a gene within an intron) were merged into a single gene prior to evaluation of merging events in order to exclude such cases from being counted as merged genes.

Results

We have compared gene prediction accuracy of GeneMark-EP, -EP+ with accuracy of GeneMark-ES. In addition, we also made an accuracy assessment of hints generated by ProtHint. We selected six species *N. crassa*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *S. lycopersicum* and *D. rerio* (Table 1). All the species but *S. lycopersicum* were model organisms with genomes expected to have sufficiently accurate annotation. Therefore, for the five model organisms we made comparisons between predicted and annotated gene co-ordinates on a whole genome scale. In case of *S. lycopersicum* we used a test set of genes validated by RNA-Seq data.

Regions of annotated pseudogenes known in genomes of *C. elegans*, *A. thaliana*, *D. melanogaster*, and *D. rerio*, were excluded from accuracy assessments. In case of *D. rerio* we excluded annotated partial exons (ubiquitous in this genome) from exon level accuracy estimations; also, the gene level sensitivity was computed only for genes with all alternative annotated transcripts being complete.

A source of protein sequences was OrthoDB v10 (23) partitioned into relevant taxonomic divisions; particularly, we used plants for *A. thaliana*, arthropods for *D. melanogaster*, etc. (as shown in Table 2).

A salient feature of the new algorithm is use of multiple homologous proteins for hints generation. In practical application, an average evolutionary distance from a protein of interest to a set of homologs could vary significantly. To model these variations in our tests, we introduced restrictions on how evolutionary close possible target proteins could be to a given query. These restrictions were implemented by removing from the database: i/ proteins originated from the species of interest per se; ii/ proteins originated from all species from the same subgenus; iii/ - from the same genus; iv/ - from the same family; v/ - from the same order; vi/ - from the same phylum. Notably, the distributions of numbers of species within the genus, family, etc. were species specific (Table 2). Given a genomic sequence and a set of reference proteins from a segment of protein database, we did run GeneMark-ES to generate seed proteins and then the ProtHint pipeline to generate whole set of mapped introns and sites along with the high-confidence subsets. Then, the remaining steps of GeneMark-EP, -EP+ were executed.

Assessment of accuracy of GeneMark-EP, -EP+

For each species (Table 1) the accuracy of ProtHint and GeneMark-EP, -EP+ was determined at gene level (Fig. 5) and exon level (Fig. S5) for several sets of reference proteins. More details on the accuracy assessment, of both GeneMark-EP (running without enforcement of high-confidence hints) and GeneMark-EP+ is given in Supplementary Materials (Table S3).

The results could be divided into three classes: fungal genomes, compact eukaryotic genomes and large eukaryotic genomes.

Gene Level Accuracy: The pattern of accuracy change at the *gene level* (Fig. 5) was similar to the one observed at exon level (Fig. S5).

Fungal genomes: *N. crassa*. Accuracy of GeneMark-ES was high, as it is typical for fungal genomes (21). GeneMark-EP+ when supported by mapped proteins from the species outside genus/order improved Sn value only by ~2% (Fig. 5a). When supported by proteins from the species outside fungal phylum the accuracy of GeneMark-EP+ matched the accuracy of GeneMark-ES (Fig. 5a). This result complemented previous observations that GeneMark-ES with unsupervised training was highly efficient *ab initio* gene finder for fungal genomes (21). We have observed earlier that addition of information from splice-aligned RNA-Seq reads for fungal genomes did not improve accuracy of GeneMark-ET in comparison with GeneMark-ES.

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) we saw an improvement by ~20% in comparison with GeneMark-ES for both Sn and Sp in *A. thaliana* and *D. melanogaster* (Figs. 5c,d). When target proteins were situated at larger evolutionary distances, the accuracy did steadily decrease. In comparison with GeneMark-ES there was an increase by 5% in gene level Sn and Sp when target proteins could be selected outside the same phylum. For *C. elegans* GeneMark-EP+ improved the accuracy of -ES by ~10% when target proteins were outside the same species (Fig. 5b); this improvement became much smaller for targets outside the same family (Sn improved by 3%, Sp unchanged). Almost no difference between -ES and -EP+ was observed if the targets were outside the same phylum. Notably, the gene level accuracy for *C. elegans* was lower than for other compact genomes.

Large eukaryotic genomes: *S. lycopersicum* and *D. rerio*. The GeneMark-ES gene level accuracy was low for large genomes (between 5% and 20%). In *S. lycopersicum*, the accuracy was improved by GeneMark-EP+ by ~15%, when protein reference sets were outside species of the same genus or order (Fig. 5e). In *D. rerio*, exclusion of proteins from the same genus or the same order led to Sn and Sp improvement by ~20% and ~5%, respectively (Fig. 5f). The improvements were twice as low when target proteins were outside the same phylum.

Lower prediction accuracy in large genomes could be partially attributed to incorrect and/or incomplete gene annotations. More detailed comparison with gene annotations in *D. rerio* and *S. lycopersicum* genomes established the following.

In *S. lycopersicum*, we observed that annotated genes supported by RNA-Seq are significantly better predicted by GeneMark-EP+ than genes without such support (Table S4). We used VARUS (29) to generate intron hints from RNA-Seq and divided annotated genes into two groups: a/ genes with all introns predicted by VARUS and b/ all other genes. GeneMark-EP+ sensitivity (with target proteins outside the *S. lycopersicum* genus) was by 40% better in set A than in set B, on gene, exon and intron levels. It is important to emphasize that RNA-Seq information was not used in GeneMark-EP+. Sensitivity measure defined for the set of introns mapped by ProtHint was also better by ~40% (Table S4).

In *D. rerio* annotation we noticed a number of partial exons that would create incomplete transcripts. We evaluated exon level Sn separately for exons within complete and incomplete transcripts (Table S5) and observed 74.6% exon Sn in complete group vs 67.5% in incomplete group. Similarly, gene level sensitivity was better by 6% in genes with complete transcripts (Table S5).

All over, we observed that for five out of six species, the accuracy of GeneMark-EP+ was better than accuracy of GeneMark-ES, regardless of the set of reference proteins used for spliced alignments (Table S3, Figs. 5, S5). Only in case of fungal genome (*N. crassa*) the improvement was negligible, the fact explained by high accuracy of an *ab initio* gene finder in fungal genomes.

Sources of improvements in gene prediction

Better performance of GeneMark-EP+ in comparison with GeneMark-ES, is expected due to i/ model parameterization on a better validated training set and ii/ enforcement of high confidence hints in gene predictions along the training process that becomes semi-supervised instead of unsupervised. Still, even when direct corrections are excluded (GeneMark-EP mode), for all the species but fungi GeneMark-EP showed improvement over GeneMark-ES. Surprisingly, GeneMark-EP showed only small fluctuations in accuracy with respect to increase in the size of the reference set of protein by including more evolutionary close species (Table S3).

The accuracy of GeneMark-EP was comparable to accuracy of GeneMark-EP+ when the smallest reference set of proteins was used (species outside the same phylum). Accuracy of GeneMark-EP+ increases more sharply than -EP when proteins from more evolutionary close species are included. The only exception was *C. elegans* in which GeneMark-EP gene level accuracy dropped by ~3% for the reference set of species outside the same phylum in comparison with GeneMark-ES (GeneMark-EP+ increased the accuracy back to the level of GeneMark-ES, Table S3).

These observations suggest that a relatively small number of anchored introns play a critical role in parameter estimation in GeneMark-EP. Further increase in the number of anchored introns does not improve parameters of GeneMark-EP. For the case of *C. elegans*, one could argue that the critical number of anchored introns was not reached when the reference set was limited to species 'outside the *C. elegans* phylum'.

To differentiate contributions into GeneMark-EP+ training, we compared use of only high-confidence intron hints with use of only high-confidence hints for gene starts and stops (Table S6**).** This experiment showed that enforceable hints of both kinds contributed equally to overall accuracy improvement. However, these hints contribute unequally into reducing different types of error. Enforcement of high-confidence intron hints led to higher prediction accuracy of internal exons, while enforcement of hints to high-confidence gene starts and stops led to reduction of errors in initial and terminal exons.

We observed that GeneMark-ES was more likely to generate gene merging than gene splitting errors (Table 4); for instance, in *A. thaliana* there were 134 split genes and 1945 merged genes. Use of GeneMark-EP (with target proteins outside the same genus) decreased frequency of errors in gene merging (a ~25% decrease in all species) however, it also caused a slight increase in gene splitting (Table 4). Transition to GeneMark-EP+ (the last column in Table 4) reduces gene merging dramatically.

Enforcement of only high-confidence intron hints reduced the number of split genes (by enforcing introns in place of incorrectly predicted intergenic regions). Still these hints have little or no effect on the gene merging (Table 4). The most significant effect was observed for *D. rerio* - 1407 split genes in the -EP+ mode compared to 2104 in the -EP mode.

Enforcement of high confidence gene start and stop hints significantly reduced the number of merged genes and caused a slight increase in the number of split genes. Number of merged genes dropped by ~1,000 in *A. thaliana* (1501 merged genes in -EP versus 453 in -EP+ with enforcing high confidence gene start/stop sites); about 50%

improvement was observed for the other species except for *C. elegans*. All over, GeneMark-EP+ (Table 4, last column) achieved significant reduction in numbers of both merged and split genes in comparison with GeneMark-ES and -EP.

Accuracy of GeneMark-EP+ vs gene annotation in the APPRIS database

Comparison of GeneMark-EP+ gene predictions with the APPRIS annotation of major protein isoforms in *C. elegans, D. melanogaster, and D. rerio* genomes (24) did show (Fig. S6) an increase in exon level sensitivity (by ~4% for *C. elegans*, by ~7% for *D. melanogaster* and *D. rerio*) and a decrease in exon level specificity (by ~1.5% for *C. elegans*, by 3% for *D. melanogaster* and by ~8% for *D. rerio*) in comparison with the accuracy assessed by comparison with the genome annotation made by a corresponding genomic community (Table 1). The decrease in Sp could be expected since the APPRIS annotation contains smaller number of exons. The increase in Sn is a positive news indicating that GeneMark-EP+ when making prediction of just one isoform per locus gets hits into genes for major protein isoforms. At gene level (Fig. S7), both Sn and Sp were reduced slightly in *C. elegans* and *D. rerio*, and by 5% in *D. melanogaster*. This result needs to be interpreted correctly in the context of conventional definition of gene level accuracy (a gene is counted as correctly predicted if the prediction matches all exons in at least one alternative transcript). Thus, prediction of one of the isoforms correctly (major or not) was all what was counted in the previously discussed results (Fig. 5, Table S3). This is a rather liberal way of computing the Sn value on gene level.

Assessment of accuracy of ProtHint

The main role of ProtHint is generation of a list of co-ordinates as well as confidence scores of potential borders between coding and non-coding regions in a novel genome. Specific thresholds on confidence scores could be define to select subsets of hints (e.g. high-confidence set). The GeneMark-EP training procedure can tolerate a high number of false positive intron hints since only a subset, the anchored introns, are used in training. It is important that the set of *all mapped hints* would have high Sn with respect to true gene elements while the Sp level could be lower. On the other hand, in the *high-confidence hints*—those utilized in initial GeneMark-EP+ parameter estimation as well as in the hints enforcement—have to have high Sp, as these hints are directly enforced in predictions.

Sensitivity of hints generated via multiple spliced alignments

When the set of reference proteins had the maximum size (all proteins outside the same species) the set of intron hints generated by ProtHint had Sn > 75% for exact introns and Sn ~70% for gene starts and stops (Tables 5, S7). The value of Sn went down steadily as the evolutionary distance to potential target proteins increased. Particularly, when the species of the same *order* were excluded, Sn was, on average, 65% for intron hints and 40% for gene start and stop hints.

The largest reduction in reference set - excluding reference proteins from the same phylum - decreased Sn of the intron hints down to 40% on average. The largest fraction of correct intron hints was observed for *N. crassa* (60%), the lowest for *C. elegans* (25%). The value of start and stop Sn generated by reference proteins from the smallest set (outside the same phylum) varied greatly between species, from ~9% for *S. lycopersicum* to ~30% for *N. crassa*. The exception to the above trend was *C. elegans*. This is explained by the fact that it had only a few relatives within the same taxonomical phylum (Table 2).

Specificity of high confidence hints generated via spliced alignments of multiple proteins

The sets of high-confidence hints were observed to have high Sp, averaging over 95% over the six species. This level remained high even for the narrowest set of reference proteins, the species outside the same phylum (Tables 5, S7). In case of *C. elegans*, along with high Sp, we observed low Sn value of high-confidence hints (in all the reference sets – larger or smaller) which could be again explained by presence of just a few species with sequenced genomes in the *C. elegans* phylum (Table 2). In all other species, the decrease in Sn in transition from all mapped to high-confidence hints was small in comparison with the simultaneous increase in Sp.

Distributions of IMC and IBA scores for introns mapped from target proteins (false and true as compared with annotation) are shown in Fig. 4a for *N. crassa* (the genus-excluded reference set). Fig. 4b shows Sn-Sp curves for filtering this set with IMC, IBA and their combination.

The distribution of score vectors shown in Fig. 4a as well as Sn-Sp curves (Fig. 4b) depend on the level of restriction on the set of reference proteins (Figure S2, left and middle panels). The IBA threshold for selection of high-confidence intron hints could affect the accuracy of GeneMark-EP+. We assessed the extent of this effect in *A. thaliana*, *N. crassa*, and *S. lycopersicum* (Figure S2, right panels). It was shown that the performance was stable with respect to changing IBA threshold; the best average prediction accuracy was achieved with IBA threshold set to 0.25. Similar effect was observed for high-confidence hints to gene starts and stops (data not shown).

More protein hints are generated in regions encoding conserved protein domains

About 50% of the whole set of introns annotated in the APPRIS set of principal isoforms was found to be located within conserved protein domains (Table S8).

In *D. melanogaster*, high-confidence introns mapped by ProtHint from the species-excluded reference set fell into regions coding for conserved domain in 55.8% of cases (Table 6). This fraction increased significantly as more proteins were excluded from the set of targets (e.g. outside the same genus) and reached 84.6% when only proteins outside the same phylum were considered (Table 6). Similar trends were observed for *C. elegans* and *D. rerio* (Table S9). In the set of *all reported introns*, the fraction of introns mapped to regions coding for conserved domains was lower than in the set of high-confidence intron hints (Table S9), however, the proportion of introns mapped into conserved domain regions also increased with removing proteins from closely related species.

The same fractions for both high-confidence and for all reported introns were almost identical between *D. rerio* and *D. melanogaster* (Table S9). For *C. elegans*, however, the figure for high-confidence introns was not close to *D. melanogaster* (Table S9) apparently due to *C. elegans* having fewer target proteins from close relatives in the protein database (the factor significantly affecting intron coverage IMC score).

Discussion

The main reason to develop GeneMark-EP, was an expectation that iterative *ab initio* parameterization of statistical models (as done in GeneMark-ES) would become more precise, especially in case of long genomes, if we find an efficient method to add data on protein footprints into training and prediction steps. This project has grown into development of a whole pipeline, still with the name GeneMark-EP. Particularly, the new pipline included ProtHint, a new method to process results of mapping of multiple homologous proteins to a genomic locus.

Currently, with millions of proteins in public databases, GeneMark-EP becomes a universal extension of GeneMark-ES, as its application to a novel eukaryotic genome will be facilitated by use of a vast volume of protein sequences.

Earlier developed GeneMark-ET (2) makes an extension of GeneMark-ES when transcriptome sequence data, with short or long reads, is available along with a newly assembled genome.

Existing methods, such as GenomeThreader (7), rely on mapping proteins from closely related species to produce accurate exon-intron structures. However, the accuracy of gene structure prediction using splice alignment of individual protein to genomic fragment is dropping significantly with increase in evolutionary distance between species (6).

Mapping multiple homologous proteins neutralizes to some degree the effect of increase of evolutionary distance. Particularly, we saw enrichment of high-confidence introns in the regions coding for conserved domains due to corroboration of hits originated from multiple homologous proteins (Table 6).

Use of anchored elements of gene structure was important for integration of signals coming from different sources (sites predicted from genomic sequence alone and sites predicteed by protein footprints). The mere principle of their selection facilitated filtering out ‘one-sided’ noise present in one or another source. Another important feature of the method was the use of partial protein footprints, when target protein mapping contributed less than full exon-intron structure. A partial contribution was still useful for improving training sets; it also could add confident corrections at the gene prediction step (Fig. S8).

Use of anchored elements was most beneficial in large genomes (*S. lycopersicum* and *D. rerio*) where GeneMark-ES alone was observed to generate a larger rate of false positive errors due to longer on average intergenic regions.

In comparison with the use of RNA-Seq reads, use of proteins allows for better discrimination between introns and intergenic regions. This occurs due to better prediction of intergenic regions with mapping of N- and C-terminals of

target proteins. We saw significant reduction of errors in gene merging (with intergenic regions predicted as introns) while error rate of gene splitting (introns predicted as intergenic regions) was less affected (Table 4).
In computational experiments for all species but fungi, *N. crassa*, we observed the most significant improvement in comparison with GeneMark-ES to occur when GeneMatk-EP+ used the largest possible sets of reference proteins (Figs. 5, S5). For *N. crassa*, use of protein evidence never helped noticeably improve the accuracy as GeneMark-ES. High accuracy of GeneMark-ES in fungal genomes was demonstrated earlier as well (21). We assume that lower performance in case of *C. elegans* in comparison with *Arabidopsis* and *Drosophila* was related to a larger number of introns per gene and lower number of reference proteins within the *C. elegans* phylum. In tomato and fish genomes that have longer on average intergenic regions than other species we saw low exon level specificity (~50-55%) related to elevated false positive prediction of protein-coding genes in long intergenic regions (Fig. S5). 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, genes in the fish genome have a rather large, 8.2, average number of introns per gene. Under independence of errors assumption, genes with large number of introns would be improbable targets for accurate prediction. Even though, the independence assumption does not hold in presence of external evidence, the pattern of gene error rate increase with the increase in number of introns was present (data not shown).
For *D. melanogaster, C. elegans* and *D. rerio* that have genome annotations in APPRIS format (24) we have shown that GeneMark-EP+ makes more accurate predictions in terms of Sn when comparison is made with the APPRIS major isoforms than when comparison is made to annotation with all possible isoforms.

Importantly, the second iteration of GeneMark-EP+ (with predictions generated by first iteration are used as the seed genes) has a small but positive effect on the final gene prediction accuracy. This additional run is recommended if there is no restriction on additional computational time.

There were several options in how to select external protein data sets as well as tools for the protein data processing. We have verified that choice of OrthoDB as a protein database, DIAMOND for search for the seed orthologs (targets) in the database and Spaln for splice alignment of targets to genome were robust with respect to outcomes of ProtHint as well as GeneMark-EP and -EP+. Choices of DIAMOND, selection of at most 25 target proteins per seed protein (Fig. S9), and Spaln were practical from the standpoint of accelerating the overall speed of the pipeline execution. We also verified that choice of GeneMark-ES for generating seeds was a faster and efficient method in comparison with the six frame translation with Procompart and ProSplign tools (8).

The discussion would be incomplete if we do not mention limitations of the new method. GeneMark-EP does not support a multiple model mode needed for genomes with heterogeneous nucleotide composition, like mammals and some plants (grasses, e.g. rice). While the current version of GeneMark-EP, -EP+ would outperform GeneMark-ES when running on such genomes, the overall accuracy could be significantly improved with more accurate modeling of genome heterogeneity.

We realize that use of taxonomic divisions for reference proteins is just the first step in accurate modeling of real-life distribution of orthologues for genes and proteins existing in a novel species. There is room for improvement for generating both intron and gene start & stop hints when reference proteins are selected based on evolutionary distance measures. Similarly, one would expect effective use of rigorous gene specific evolutionary distance in selecting thresholds for intron mapping.

Another limitation of the method is the search for a single optimal solution that leads to prediction of a single gene, single protein isoform in each locus. Importance of genes with alternative splicing has been debated recently, as the evidence was accumulated that alternative splicing mainly operates with UTR regions rather than with translated regions of pre-mRNA. Moreover, the claims were made that when a translated region could be alternativly spliced then only one among the protein isoforms, the major one, is expressed in significantly large number of copies than the minor ones. If gene prediction by GeneMark-EP, -EP+ is viewed as prediction of the major isoform, then the result should be naturally assessed in comparison with annotation of the major isoforms. Such annotation is provided by the APPRIS database and the comparison was done for *C. elegans, D. melanogaster, and D. rerio*. Nonetheless, general tools that use external information to predict alternative isoforms are of significant interest for community. Particularly interesting case is when external information representing alternative isoforms at RNA level. In this case, a pipeline, BRAKER1 (30) makes predictions of alternative isoforms by GeneMark-ET and AUGUSTUS. A new pipeline, BRAKER2 (paper in preparation) combines GeneMark-EP, -EP+ with AUGUSTUS to identify a set of alternative protein isoforms when variants of cross-species proteins are

given among references. A new tool, GeneMark-ETP, will use protein, and transcript data available for each locus (paper in preparation).

Availability

GeneMark-EP+, ProtHint, and all scripts and data used to generate figures and tables in this manuscript are available at https://github.com/gatech-genemark/GeneMark-EP-plus. Software is compiled for Linux 64 bit operating system. To give an example, the overall runtime of ProtHint and GeneMark-EP+ on *D. melanogaster* genome (of 134 MB and ~14,000 genes) with target proteins from species outside *Drosophilidae* family was ~5 hours on an 8CPU/8GB RAM machine. In our experiments the run time grew linearly with genome length and number of genes.

Funding

This work was supported in part by the National Institutes of Health (NIH) [GM128145 to M.B.].
Funding for open access charge: National Institutes of Health [GM128145].

Conflict of interest statement

None declared.

References

1. Hoff, K.J. and Stanke, M. (2019) Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics*, **65**, e57.
2. Lomsadze, A., Burns, P.D. and Borodovsky, M. (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*, **42**, e119.
3. Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P. and Schiex, T. (2008) Genome annotation in plants and fungi: EuGene as a model platform. *Curr Bioinform*, **3**, 87-97.
4. Sallet, E., Gouzy, J. and Schiex, T. (2019) EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. *Methods Mol Biol*, **1962**, 97-120.
5. Behr, J., Bohnert, R., Zeller, G., Schweikert, G., Hartmann, L. and Rätsch, G. (2010) Next generation genome annotation with mGene.ngs. *BMC bioinformatics*, **11**, O8.
6. Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. *Genome Res*, **14**, 988-995.
7. Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inform Software Tech*, **47**, 965-978.
8. Kiryutin, B., Souvorov, A. and Tatusova, T. (2007), *11th Annual International Conference in Research in Computational Molecular Biology*, San Francisco, USA.
9. Gotoh, O. (2008) Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics*, **24**, 2438-2444.
10. Keller, O., Kollmar, M., Stanke, M. and Waack, S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757-763.
11. Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J. and Hartung, F. (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, **44**.
12. Gotoh, O., Morita, M. and Nelson, D.R. (2014) Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *Bmc Bioinformatics*, **15**, 189.
13. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, **268**, 78-94.
14. Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, **26**, 1107-1115.
15. Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19 Suppl 2**, ii215-225.
16. Parra, G., Blanco, E. and Guigo, R. (2000) GeneID in Drosophila. *Genome Res*, **10**, 511-515.
17. Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T. and Lipman, D. (2010) Gnomon–NCBI eukaryotic gene prediction tool.
18. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.

19. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Giron, C., Hourlier, T. *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)*, **2016**.
20. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**, 6494-6506.
21. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research*, **18**, 1979-1990.
22. Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061-1067.
23. Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F.A. and Zdobnov, E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*, **47**, D807-D811.
24. Rodriguez, J.M., Rodriguez-Rivas, J., Di Domenico, T., Vazquez, J., Valencia, A. and Tress, M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res*, **46**, D213-D217.
25. Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, **12**, 59-60.
26. Borodovsky, M. and Mcininch, J. (1993) Genmark - Parallel Gene Recognition for Both DNA Strands. *Comput Chem*, **17**, 123-133.
27. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *Bmc Bioinformatics*, **10**, 421.
28. Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R. *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*, **45**, D200-D203.
29. Stanke, M., Bruhn, W., Becker, F. and Hoff, K.J. (2019) VARUS: sampling complementary RNA reads from the sequence read archive. *Bmc Bioinformatics*, **20**, 558.
30. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767-769.

Figures

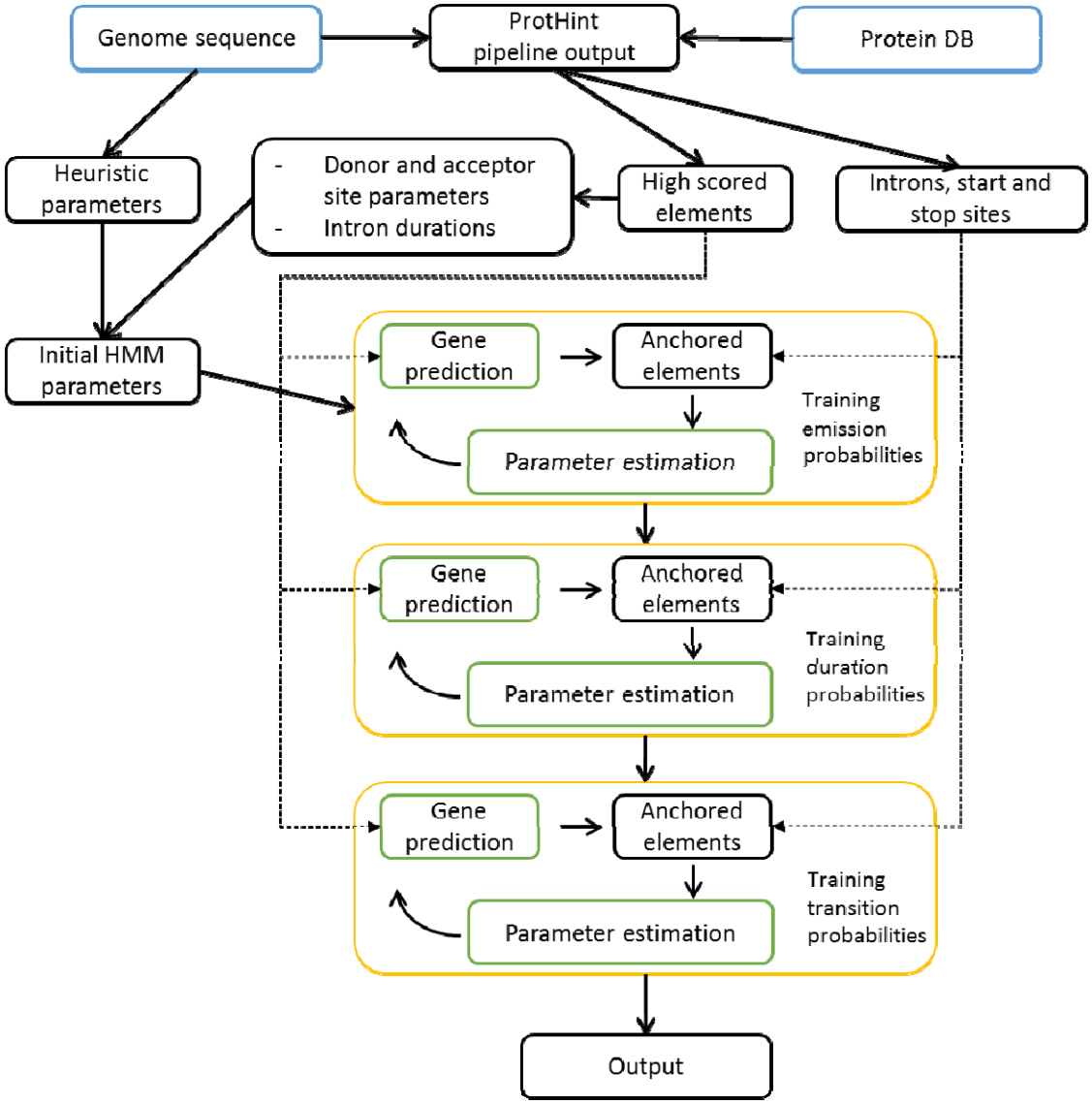

Figure 1: GeneMark-EP training diagram.

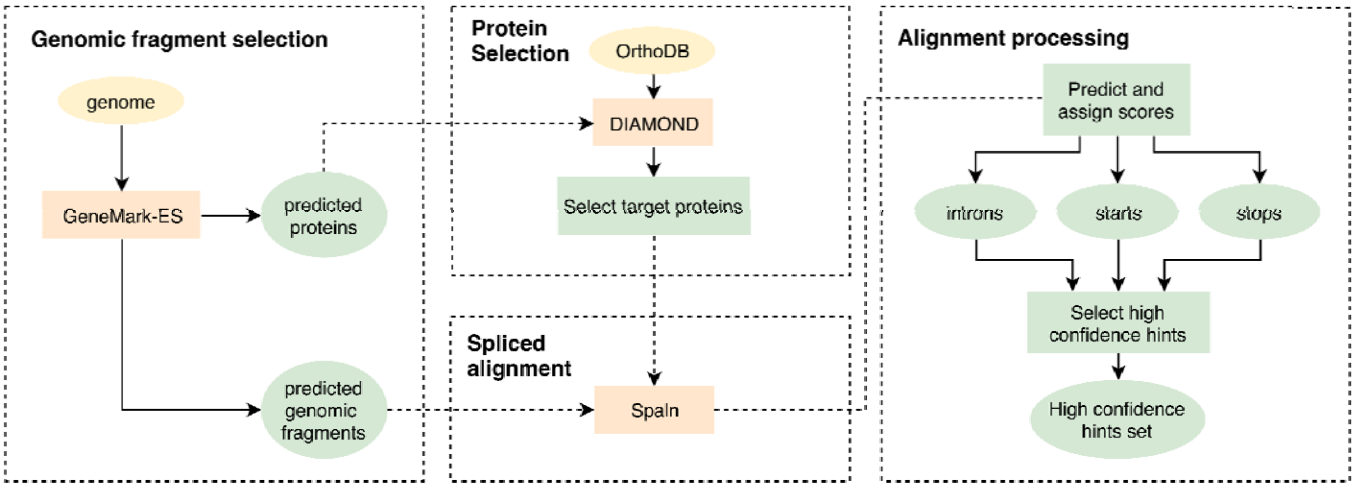

Figure 2: An overview of the ProtHint pipeline.

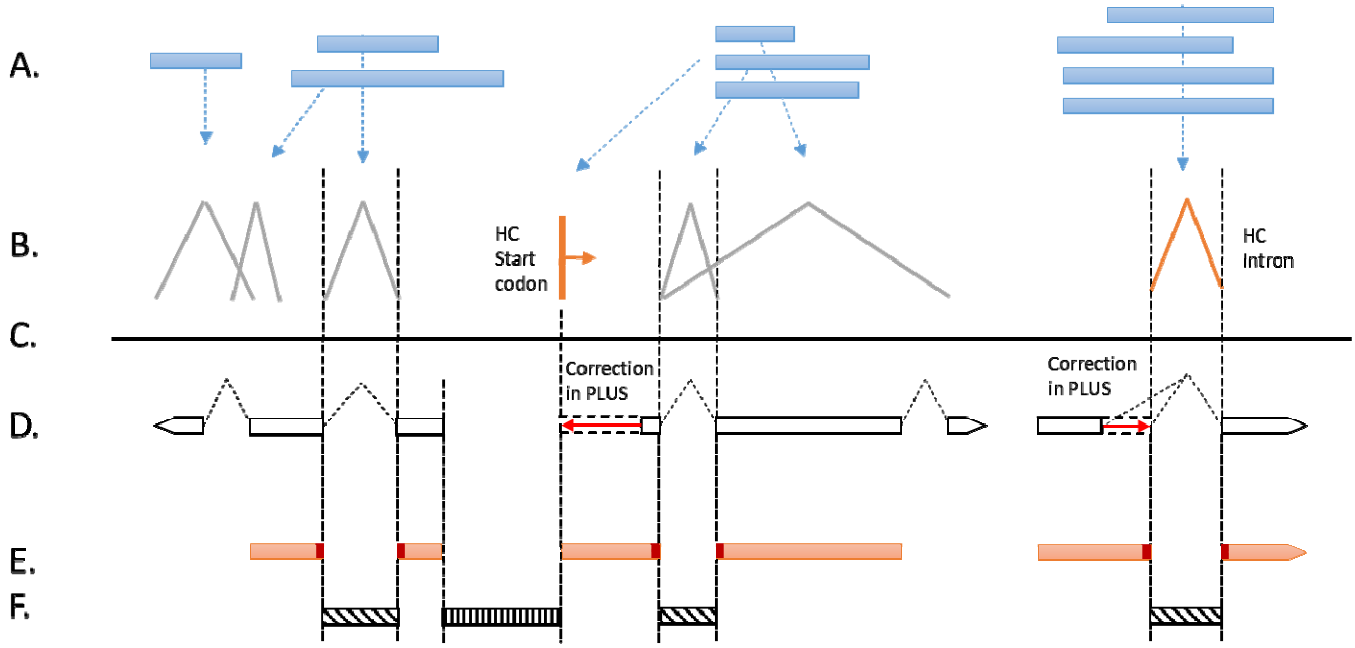

Figure 3: Selection of sequence regions for GeneMark-EP+ training with enforcement of High-Confidence (HC) hints.

A. Target proteins
B. Introns, start and stop sites defined by spliced alignments of target proteins to genome
C. Genome sequence
D. Gene prediction by GeneMark.hmm at a given iteration
E. Selection of splice sites and protein coding regions for training (anchored elements)
F. Selection of non-coding regions for training

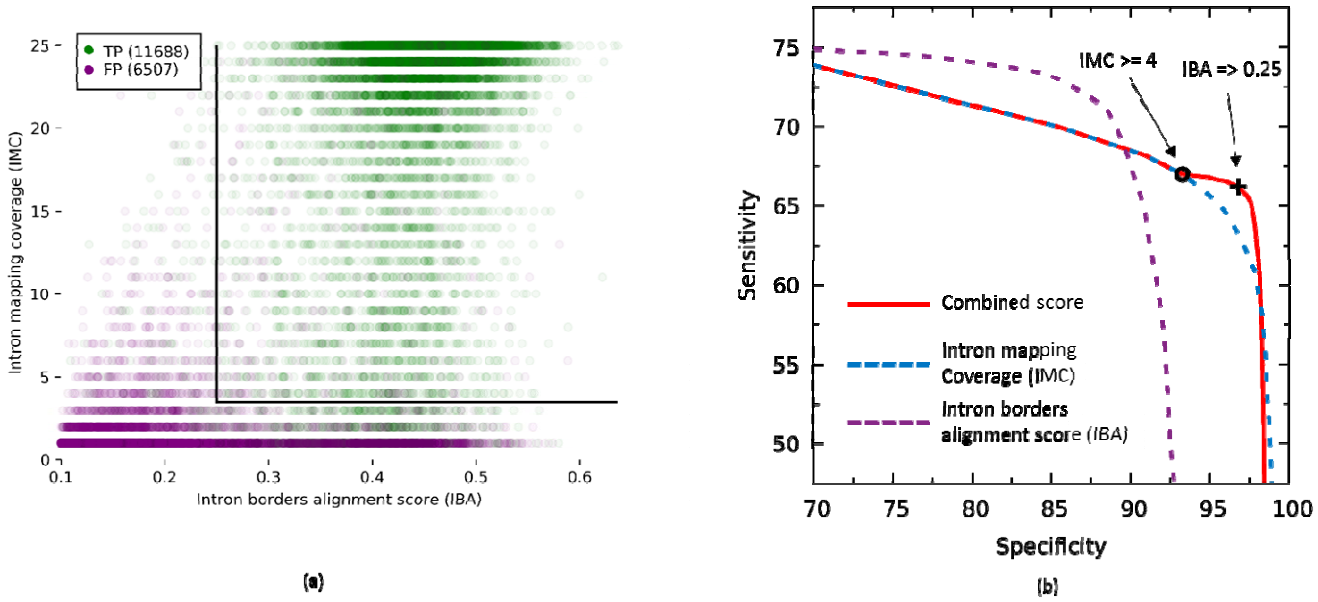

Figure 4: ProtHint mapping of introns in case of *N. crassa*. Mapping is done from target proteins that belong to species beyond *Neurospora* genus (a) Distribution of the score vectors (IBA, IMC) of true positive (green) and false positive (purple) introns mapped by spliced alignments. The black lines 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 intron borders alignment (IBA) score and intron mapping coverage (IMC) score. IMC is computed for introns which have IBA score $\geq$ 0.1 and exon AEE score $\geq$ 25. Red Sn-Sp curve represents a combination of scores: The curve is generated by first, selecting all introns above IMC threshold changing from 0 to 4 and then selecting all the introns with IBA score changing from 0 to 0.25 and up to 1.0. The crossed position in the red curve represents IMC $\geq$ 4 and IBA $\geq$ 0.25 thresholds. Separate curves for IMC score (dashed blue) and IBA score (dashed purple) are shown as well.

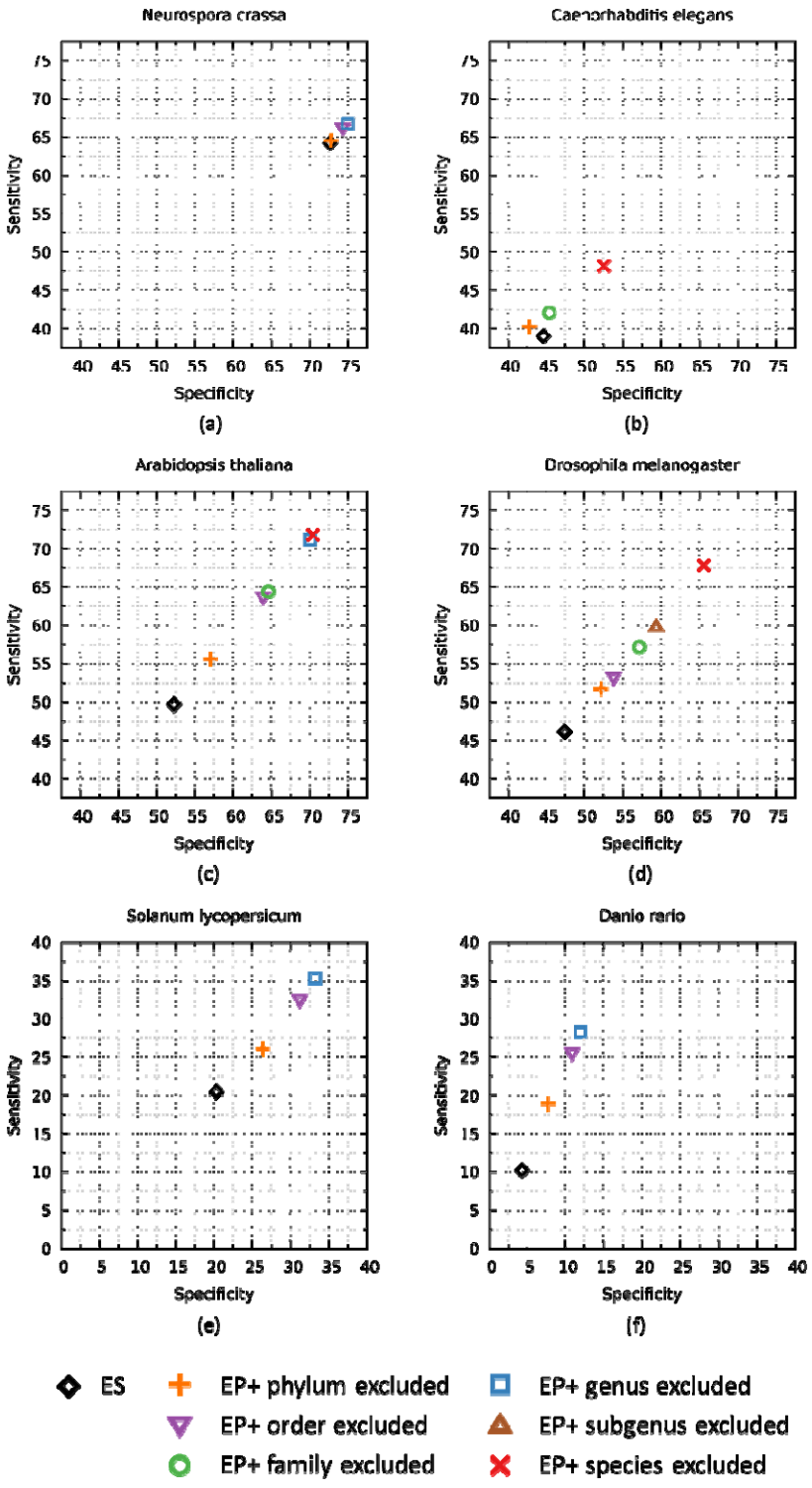

Figure 5: Comparison of GeneMark-ES and GeneMark-EP+ accuracy on gene level accuracy of GeneMark-EP+ is shown for cases when ProtHint works with different sets of reference OrthoDB proteins: from the largest (only the same species excluded) to the smallest (the whole same phylum excluded). A gene prediction is considered to be correct if it matches one of the annotated isoforms. Gene level Sn of *D. rerio* was computed only with respect to complete genes.

Tables

Species	Assembly version (NCBI)	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	GCF_000188115	807	Consortium ITAG3.2 (Jun. 2017)	34,950	3.7
Danio rerio	GCF_000002035	1,345	Ensembl GRCz11.96 (May 2019)	25,254	8.2

Table 1: Genomes used for tests of GeneMark-EP and GeneMark-EP+. Introns per gene are computed for all genes, including single-exon genes.

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	Arthopoda	2,601,995
Solanum lycopersicum	1	9	10	-	99	**116**	Plantae	3,510,742
*Danio rerio**	0	4	4	49	**245**	447	Chordata	5,003,104

Table 2: Characteristics of the OrthoDB v10 taxonomical space for each of the species. The size of the largest sections of the database (OrthoDB root) used for each species is shown in bold.
*For tests in the phylum-excluded mode, the kingdom was used as the root.

		All reported starts	Filtered with SMC >= 4	Filtered with SMC >= 4 and exon overlap =0
A. thaliana	Sn	67.6	61.0	59.6
	Sp	72.1	90.3	**94.7**

Table 3: ProtHint sensitivity and specificity of all predicted and of the high-confidence starts. High specificity is achieved with filtering by site mapping coverage (SMC) scores as well as by removal of candidate starts overlapped by at least one target protein which suggests an alternative start upstream. Sensitivity is defined with respect to the full complement of starts, including alternative ones. The numbers were generated in the test in genus-excluded mode. Results for all test species are shown in supplementary table S2.

	Genes	ES	EP	EP+ Introns (a)	EP+ Starts/Stops (b)	EP+ Full (c)
N. crassa	Merged	132	94	97	70	78
	Split	73	85	82	92	85
C. elegans	Merged	2335	2047	2043	1896	1893
	Split	283	396	351	405	355
A. thaliana	Merged	1945	1501	1445	453	475
	Split	134	175	137	201	145
D. melanogaster	Merged	945	759	742	513	514
	Split	143	152	107	168	123
S. lycopersicum	Merged	3213	2643	2572	1724	1753
	Split	688	820	626	950	740
D.rerio	Merged	2603	1845	1749	1150	1143
	Split	1667	2104	1407	2053	1383

Table 4: Numbers of merged and split genes in predictions of GeneMark-ES, -EP and -EP+ with enforcement of (a) high confidence introns only, (b) high confidence starts and stops only (c) enforcement of both (a) and (b). All the numbers were obtained for reference sets of target proteins defined for the genus-excluded mode.

	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	78.7	73.3	71.9	61.6	65.4	53.4	48.9	33.8	35.2	20.5
Intron Sp	83.8	**98.9**	79.8	**98.9**	79.6	**98.8**	80.5	**99.0**	88.4	**99.5**
Start Sn	69.1	59.7	49.0	35.9	37.1	28.8	21.9	15.7	13.9	9.5
Start Sp	80.3	**97.5**	76.6	**96.9**	72.4	**95.9**	74.3	**94.6**	75.8	**93.8**
Stop Sn	74.0	67.0	55.7	44.3	43.8	36.2	26.1	19.4	15.5	11.0
Stop Sp	95.3	**99.3**	94.6	**98.9**	93.2	**98.6**	94.8	**99.0**	96.2	**99.3**

Table 5: ProtHint protein mapping performance for *D. melanogaster*: Sensitivity and specificity of detection of introns, start and stop codons. The results are shown for the use of *all reported* hints or just *high-confidence* hints. The accuracy is computed based on genome annotation. The accuracy is defined with respect to the full complement of introns, starts and stops, including alternative splicing. Results for all test species are shown in supplementary table S7.

Exclusion level	High-confidence introns matching APPRIS introns		
	All	In domains	
Species	33,334	18,616	(55.8%)
Subgenus	27,999	17,190	(61.4%)
Family	24,274	15,791	(65.1%)
Order	15,538	11,768	(75.7%)
Phylum	9,519	8,057	(84.6%)

Table 6: Statistics for *D. melanogaster* genome. Fraction of high-confidence intron hints mapped into regions coding for conserved protein domains for different sets of reference proteins. Out of 41,010 introns in the APPRIS *D. melanogaster* genome annotation, 21,562 (52.6%) are located in regions encoding conserved protein domains.