Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

Current methods for automated annotation of protein-coding genes

KJ Hoff and M Stanke

We review software tools for gene prediction — the identification of protein-coding genes and their structure in genome sequences. The discussed approaches include methods based on RNA-Seq and current methods based on homology — comparative gene prediction and protein spliced alignments. Many methods require that their parameters are adjusted to the target species or its broader clade. These include *ab initio* gene finders, integrated approaches with *ab initio* components and some aligners. We also review current automatic methods for training for the common case that a *bona fide* training set of gene structures is not available before annotation.

Address
Institut für Mathematik und Informatik, Universität Greifswald, Walther-Rathenau-Str. 47, 17487 Greifswald, Germany

Current Opinion in Insect Science 2015, **7**:8–14

This review comes from a themed issue on **Insect genomics**

Edited by **Susan Brown** and **Denis Tagu**

For a complete overview see the Issue and the Editorial

Available online 7th March 2015

http://dx.doi.org/10.1016/j.cois.2015.02.008

Introduction

The accurate structural annotation of protein-coding genes is an early and important step in the analysis of assembled genomes because further downstream analysis such as the study of protein family evolution [1] and the experimental investigation of selected genes may be misguided or may fail with a structural annotation of low quality. Software tools that utilize statistical models to predict protein-coding genes in genomic sequences are now often called *ab initio* methods, although in the original more strict sense *ab initio* refers to methods that use no evidence but the target genome itself. Commonly used additional evidence comes from RNA-Seq or expressed sequence tags (ESTs), from protein databases, from mass spectrometry or from the genomes of related species. Such evidence helps to improve the accuracy of the genes or gene parts for which it is available. In this short review, the prediction of protein-coding genes is covered. For other aspects of genome annotation, such as the annotation of repeats, pseudogenes and noncoding RNAs we refer the reader to the extensive review of Haas *et al.* [2].

Many methods and pipelines for gene finding in eukaryotes are universally applicable. That does not imply that the accuracy or even relative accuracy of tools is transferable between species. Intron size has a large influence on the difficulty and accuracy of gene prediction. The presence of long introns in a species makes errors such as false positive exons and the splitting of a gene into several predicted ones or the joining of several genes to one predicted gene more likely. In addition, genomes with long introns allow for more complex alternative splicing.

Gene prediction in insects typically takes an intermediate place in terms of the difficulty of the task and the accuracy of methods between the more difficult vertebrates and the less difficult fungi and algae. In insects, introns may typically be long enough to allow for complex alternative splicing and false predicted exons within the range of a true intron. On the other hand, the median insect genome size is about 6-fold smaller than the median vertebrate genome size, which makes it easier to achieve a good specificity.

Training of gene finders

Ab initio prediction is required for those genes that are weakly or not at all represented in any RNA-Seq library, that have insufficient similarity to any known protein and lack other evidence [3]. In addition, *ab initio* components are used to identify the protein translation of a transcript with otherwise known exon–intron structure or to identify the structure of the non-conserved parts of a gene with partial homology information. Further, *ab initio* components can help to choose correct splicing structures in the common case that transcriptome assembly or protein spliced alignment allow many possible gene structures.

Ab initio methods 'learn' specifics of protein-coding gene structures in the target genome, like splice site patterns, the translation start and other biological signals, typical sequence composition and intron and exon length distributions. However, most *ab initio* gene prediction tools need a set of at least a few hundred initial example genes to train the parameters (e.g. AUGUSTUS [4], SNAP [5], GeneID [6]).

The necessity to train and the possibilities for the training of a gene finder depend on the phylogenetic proximity to

other well-annotated species or species with pre-trained parameters, and on the availability of transcriptome and other experimental data. A fairly generally applicable method is the prediction of the structure of a subset of core eukaryotic proteins via the CEGMA pipeline [7]. Homolog proteins highly similar to a gene in the target genome can also serve as a source to build initial genes with Scipio [8]. Alternatively or additionally, if EST or RNA-Seq data is available, the PASA pipeline [9] and MAKER2 [10] produce training gene structures.

Recently, fully automatic training methods became available. WebAUGUSTUS [11•] is a web server implementation of the autoAug pipeline of the AUGUSTUS distribution. Among others, it trains AUGUSTUS from sequence input alone (protein or transcript sequences, employing PASA or Scipio) and may also be used to predict genes genome-wide. GeneMark-ET automates the training of the *ab initio* gene finder GeneMark from RNA-Seq [12•]. The sister tool GeneMark-ES requires only the genome itself for training [13] but may be less accurate. SnowyOwl [14] is a training and annotation pipeline that was tested on fungi and that combines initial transcript models from assembled RNA-Seq data and GeneMark-ES to train AUGUSTUS, subsequently, genes are predicted with AUGUSTUS and RNA-Seq hints and the resulting gene models are combined with GeneMark-ES predictions.

Transcript-based approaches

Genome sequencing is now usually accompanied by large scale transcriptome sequencing, mostly RNA-Seq. There are three major types of approaches for transcriptome-based gene finding, depicted at the top left of Figure 1. All approaches require the spliced alignment of transcript sequences — either single reads or assemblies thereof — to a genome of the same or closely related species. Suitable tools for the spliced alignment of RNA-Seq reads are for example STAR [15], GSNAP [16], TopHat2 [17] and PALMapper [18]. The alignments provide information about the location and structure of transcripts, that is introns and exons, and on transcript alternatives such as alternative splicing or alternative transcription initiation or termination. Transcriptome assemblers like MITIE [19], Cufflinks [20] and StringTie [21] construct from the alignments a set of transcripts for each locus but do not predict whether they encode a protein. This can be done in an additional step (Figure 1).

In 2013, results of the RNA-Seq Genome Annotation Assessment Project (RGASP) were published [22••], independently assessing results of 14 transcriptome reconstruction and gene prediction methods on human, *D. melanogaster* and *C. elegans* submitted by the tool authors themselves in 2010. RGASP was open to submissions from the field of *de novo* transcript reconstruction and methods that infer transcript structures from read alignments to the genome; six of the participating tools predicted coding sequences in the transcripts (e.g. AUGUSTUS [23], mGene [24,25], Transomics [26]), the others, for example Cufflinks [20], did not. On *Drosophila melanogaster*, the best performing tools for protein-coding gene prediction were AUGUSTUS (48.53%/44.03%), Transomics (46.95%/33.54%), and mGene (43.99%/44.02%). These numbers refer to sensitivity and precision on the gene level. For example, for 48.53% of the genes in the FlyBase reference gene set, AUGUSTUS predicted at least one of its protein isoforms exactly, that is without any errors. A more stringent evaluation criterion is the percentage of the reference protein isoforms that were predicted correctly, including all reference alternative transcripts. Here, the maximum achieved value was only about 24% on *Drosophila* and about 20% on human (achieved by AUGUSTUS and exonerate, respectively). In his comment, Korf calls the RGASP results 'a little depressing' because of the disappointingly low accuracy [27] and concludes that the methods that contain a model of gene structure (AUGUSTUS, mGENE, Transomics) perform better because they know what genes are supposed to look like.

Although improvements in annotation quality are to be expected from improvements of sequencing technology, of alignment programs and of gene-finders, we are not aware of studies that report a dramatically improved accuracy of genome-wide RNA-Seq-based protein-coding gene prediction over the results from RGASP. Nevertheless, RNA-Seq is a valuable information resource for structural genome annotation. Still, many open challenges in the utilization of this data remain: Transcript sequences do not provide evidence for translation or, if a transcript is translated, which open reading frames are translated. The prediction of protein-coding genes is hindered by the presence of random non-translated ORFs in the transcribed sequences. This may include long untranslated regions (UTRs), incompletely processed RNAs (retained, possibly long introns) and noncoding genes (Figure 1).

Homology-based approaches

The protein sequence and the exon–intron structure of a gene can be fairly conserved through wide branches of the tree of life. For example, Csuros *et al.* report that introns have mostly been lost since the intron-dense most recent common ancestor of all Metazoans, more so in insects than in mammals [28], which suggests that many of the introns in insects are also conserved in mammals. The information from homology is exploited by current gene prediction methods in mainly two ways, through *protein spliced alignments* and through *comparative gene prediction*.

Protein spliced alignment methods use as input either a single protein sequence (e.g. exonerate [29], Spaln [30•], ProSplign [31]) or a representation of a protein family