mGene.web: a web service for accurate computational gene finding

Citation for published version:

Digital Object Identifier (DOI):
10.1093/nar/gkp479

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
Nucleic Acids Research

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
mGene.web: a web service for accurate computational gene finding

Gabriele Schweikert1,2,3, Jonas Behr1, Alexander Zien1,4, Georg Zeller1,3, Cheng Soon Ong1,2, Sören Sonnenburg1 and Gunnar Rätsch1,*

1Friedrich Miescher Laboratory, Max Planck Society, 2Department for Empirical Inference, Max Planck Institute for Biological Cybernetics, 3Department for Molecular Biology, Max Planck Institute for Developmental Biology Tübingen, Germany and 4Department for Intelligent Data Analysis, Fraunhofer Institute FIRST, Berlin, Germany

Received March 5, 2009; Revised May 7, 2009; Accepted May 17, 2009

ABSTRACT

We describe mGene.web, a web service for the genome-wide prediction of protein coding genes from eukaryotic DNA sequences. It offers pre-trained models for the recognition of gene structures including untranslated regions in an increasing number of organisms. With mGene.web, users have the additional possibility to train the system with their own data for other organisms on the push of a button, a functionality that will greatly accelerate the annotation of newly sequenced genomes. The system is built in a highly modular way, such that individual components of the framework, like the promoter prediction tool or the splice site predictor, can be used autonomously. The underlying gene finding system mGene is based on discriminative machine learning techniques and its high accuracy has been demonstrated in an international competition on nematode genomes. mGene.web is available at http://mgene.org/web, it is free of charge and can be used for eukaryotic genomes of small to moderate size (several hundred Mbp).

INTRODUCTION

In recent years, the biological community has started to see the dramatic impact of new sequencing technologies on the number of sequenced genomes, and it is expected that this influx of data will continue to escalate in the near future. The demand for efficient, highly automated DNA sequence analysis tools is therefore greater than ever. In particular, we expect that the task of genome annotation will increasingly be performed by individual labs rather than large sequencing centers with dedicated resources and specialized expertise. One of the most important sub-tasks in such an annotation pipeline is the identification of protein coding genes. It requires a computational gene finding system that (i) is highly accurate, (ii) produces genome-wide predictions within a reasonable time, (iii) is easy to use even for researchers with no programming experience and (iv) is applicable to a large variety of newly sequenced organisms. Since computational gene finding has a long tradition in bioinformatics research, there have been constant advancements toward this goal. In particular, the accuracy of computational gene finding systems has steadily been improved, most recently by the introduction of discriminative machine learning techniques (1). However, for this new generation of algorithms, such as mSplicer (14), Craig (2), Conrad (3), Contrast (4), and mGene (Schweikert et al., under review), there is no easy-to-use web application available. To employ these tools, the respective packages have to be downloaded and installed, which in some cases requires substantial programming knowledge as well as the accessibility of sufficient computational power for each user. On the other hand, some of the conventional state-of-the-art gene finders that use generative models, such as Fgenesh (5) and Augustus (6) offer web services, however, without the functionality of training on data provided by the user. Yet, this is essential for accurate annotation of newly sequenced organisms, as models trained on sequences derived from other organisms may produce highly incorrect predictions (7). In 2008, Ter-Hovhannisyan et al. introduced a new ab initio algorithm, GeneMark-ES, that performs unsupervised self-training on anonymous eukaryotic sequences (8). Currently, there is no web service available for GeneMark-ES, in contrast to their self-training prediction program for prokaryotic genomes, GeneMark-S (9). Therefore, to the best of our knowledge, to date no system completely satisfies all the

*To whom correspondence should be addressed. Tel: +49 7071 601 820; Fax: +49 7071 601 801; Email: gunnar.raetsch@tuebingen.mpg.de

Present address: Cheng Soon Ong, Department of Computer Science, ETH, Zürich, Switzerland.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
The precise method for computing accurate gene segmentation, thereby ultimately leading to more accurate improvements in the complex process of computational feature may guide the systematic exploration of further sub-task become available. We expect that this particular individual units can also easily be replaced by the user themselves carefully crafted and highly accurate (12,13).

transcription starts or splice sites. These predictors are independently to predict signals on the DNA, for example mGene.web, the individual sub-tools can be employed omes. Furthermore, due to the modular structure of mutation databases as well as other computational tools. We provide six signal sensors for the detection of trancription start and stop sites, translation initiation and which also offers handy access to existing genome anno- tation databases as well as other computational tools. In contrast to most other systems, mGene.web easily allows users to train the prediction model for new genomes. Furthermore, due to the modular structure of mGene.web, the individual sub-tools can be employed independently to predict signals on the DNA, for example transcription starts or splice sites. These predictors are themselves carefully crafted and highly accurate (12,13). However, in combination with Galaxy workflows, the individual units can also easily be replaced by the user when other, possibly more advanced tools for a given sub-task become available. We expect that this particular feature may guide the systematic exploration of further improvements in the complex process of computational gene finding, thereby ultimately leading to more accurate gene predictions.

**METHODS: mGene**

The precise method for computing accurate gene segmentations employed in mGene is described in detail in (G. Schweikert, under review). Here, we only give a short sketch of the underlying technique, which will help to understand how mGene.web works.

Gene finding can be viewed as a segmentation task, where a piece of genomic DNA has to be properly segmented into genic components, e.g. intergenic regions, UTRs, exons and introns. Such a segmentation is characterized by signals on the genomic DNA that demarcate individual segment boundaries, for instance, splice sites that delineate introns. Each segment type further exhibits a characteristic sequence content, e.g. nucleotide triplets characteristic of coding sequence.

We have taken a two-layered approach and subdivided the problem of gene finding into several sub-tasks that can be solved independently of each other. These tasks include the detection of sequence signals and segment contents. We provide six signal sensors for the detection of transcription start and stop sites, translation initiation and termination sites, as well as donor and acceptor splice sites (12–15). In addition to the signal sensors, we have devised content sensors, which use the sequence of complete segments as input to predict the segment type, i.e. intergenic, 5'-UTR, coding exon, intron or 3'-UTR. Each of these prediction problems is approached with a large-scale implementation of support vector machines using string kernels (as implemented in the Shogun toolbox) that are specifically designed to accurately identify signals and content types [16, and references therein].

In the second layer, the individual signal and content predictions are reconciled according to a model for gene structures (Figure 1), thereby solving the segmentation task. For this we employ hidden semi-Markov support vector machines (HSM-SVMs) (17), an approach that is conceptually similar to generalized hidden Markov models (gHMMs) (18) in that they are both based on a state-transition model which explicitly parameterizes segment lengths. However, in contrast to generative gHMMs, HSM-SVMs are trained discriminatively. For prediction we use dynamic programming to determine the highest scoring segmentation according to the trained model parameters, thereby predicting accurate gene structures even for new sequences [see for details (15,17) and (G. Schweikert, under review)].

### Table 1. Comparison of the top-performing gene finding systems in the ab initio setting of the nGASP challenge (10)

<table>
<thead>
<tr>
<th>Method</th>
<th>Nucleotide</th>
<th>Exon</th>
<th>Transcript</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>Sp</td>
<td>Sn+Sp</td>
<td>Sn</td>
</tr>
<tr>
<td>mGene.init</td>
<td>96.8</td>
<td>90.9</td>
<td>93.8</td>
<td>85.1</td>
</tr>
<tr>
<td>mGene.init (dev)</td>
<td>96.9</td>
<td>91.6</td>
<td><strong>94.2</strong></td>
<td>84.2</td>
</tr>
<tr>
<td>Craig</td>
<td>95.5</td>
<td>90.9</td>
<td>93.2</td>
<td>80.3</td>
</tr>
<tr>
<td>Fgeneshe</td>
<td>98.2</td>
<td>87.1</td>
<td>92.7</td>
<td>86.4</td>
</tr>
<tr>
<td>Augustus</td>
<td>97.0</td>
<td>89.0</td>
<td>93.0</td>
<td>86.1</td>
</tr>
</tbody>
</table>

Shown are sensitivity (Sn), specificity (Sp) and their average (each in percent) on nucleotide, exon, transcript and gene levels (if several submissions were made for one method, we chose the version with the best gene level average of sensitivity and specificity). The predictions of mGene.init were prepared after the deadline but strictly adhering to the rules and conditions of the nGASP challenge. The result of the best-performing method according to each of the evaluation levels is set in bold face. The evaluation is based on the submitted sets of the participants and performed with our own routine. The numbers slightly deviate from the official nGASP evaluation on the transcript and gene level due to minor differences in the evaluation criteria. These differences, however, do not change the ranking.
WEB SERVER

Essentially, there are two possibilities to use mGene.web for gene predictions: the simplest way is to use the monolithic tools mGenePredict and mGeneTrain. However, for a more flexible use and a more detailed output, one can also use the individual modules together with workflows. In this case, the user has the advantage of accessing intermediate results of the gene prediction task; for example, the outputs of the signal and content sensors can be monitored. Additionally, the progress can be easily observed as the individual tasks are completed. If only sub-tasks should be performed, e.g. the prediction of splice sites, the individual tools can be applied independently. Each option will be described below, and more instructions are provided at http://mgene.web/examples.

mGene.web monoliths

For organisms that are already in the list of pre-trained models, the simplest way to annotate a genome sequence with mGene.web is to use mGenePredict. This function only requires a FASTA file with the DNA sequence as input and it outputs the GFF3 file containing gene predictions. In case one wishes to annotate a genome for which no suitable pre-trained model exists, mGeneTrain can be used to train a new model. This tool takes a FASTA file with the DNA and a GFF3 file with a set of known genes as input and returns a trained mGene predictor (TmGP) object that can be used together with mGenePredict in order to predict genes on given DNA sequences. The predictor and the predictions can be easily shared with other users via Galaxy’s ‘share history’ functionality. We also offer the tool mGeneEval to compare two GFF3 files, presumable one containing the ground truth (an annotation file) and another one containing the predictions. The resulting performance report includes sensitivity and specificity values on nucleotide, exon, transcript and gene level as well as for several signals (like transcription start and stop site) and enables a convenient quality control. For instructions and example runs go to mGene.web → Examples and Instructions and http://mgene.org/web/examples.

mGene.web workflows

Recently, Galaxy has started to support workflows that allow pre-defining the order in which tools are applied to data files (including original inputs or interim results) in order to achieve a certain goal. These workflows can be edited by, and shared among users. Workflows can help to simplify the relatively complex process of creating a gene finding system which involves many different steps. In order to run a workflow, one only needs to specify a few input arguments. As a result one obtains, for instance, the gene predictions as well as intermediate results allowing detailed inspection. Analogous to the simple black-box versions, we provide two workflows (The monolithic workflows described above are equivalent to the modular versions. They show less internal details and are slightly more efficient): mGenePredict and mGeneTrain. For more information go to mGene.web workflows → Examples and Instructions or http://mgene.org/web/workflows. There, you will also find instructions on how to built new workflows, in which individual tools are replaced by other, potentially improved ones.

mGene.web modules

The web service currently provides 14 core modules. They can be grouped into four groups: data preparation; signal training and prediction; content training and prediction; and gene structure training and prediction. Each tool requires a set of inputs and provides at least one output. They are managed by the Galaxy system according to their data types. We use the following data types for data exchange between the modules: FASTA (genomic DNA), GFF3 (genome annotation and gene prediction), SPF (signal prediction format) and CPF (content prediction format). Files in these formats can either be provided by uploading or by running one of the tools. Additionally, we have some data formats representing internal data
structures, including the Genome Information Object (GIO) to describe a genome-wide sequence with all its contigs or chromosomes, the Annotation Gene Structure (AGS), as an efficient internal representation of genes, the Trained Signal Predictor (TSP), the Trained Content Predictor (TCP) and the Trained Gene Predictor (TGP) that include all learnt parameters necessary to predict a given signal, content or gene structure, respectively. Eventually, the trained mGene predictor (TmGP) contains all above-mentioned parameters and can be used to predict genes from scratch when given a DNA sequence. (Note that the TGP object contains the parameters learnt during training layer 2 only, while the TmGP contains all parameters necessary to predict genes.) In the following, the individual modules are explained in more detail (see also mGene.web modules → Examples and Instructions and http://mgene.org/web/modules).

Data preparation
GenomeTool needs a file in FASTA format containing genomic sequences as input that allows it to create a genome object, stored in a GIO, to be used by other mGene modules. Additionally, one may create a GIO from an internal database of more than 50 genomes. GFF2Anno reads an annotation file in GFF3 or GTF format and uses it to generate an AGS. It provides pre-defined settings for commonly used GFF-encoding conventions, but also permits specific settings needed for less common encoding choices.

Signal prediction
Anno2SignalLabel uses an AGS to collect labeled genomic positions for the selected genomic signal. Possible signals include transcription start and stop sites, translation initiation and termination sites, as well as donor and acceptor splice sites. It uses the regions covered by annotated features to generate negative examples at all consensus positions unless they were annotated as true sites. The output is a file in SPF providing chromosome/contig name, position, strand and the label of the example.

SignalTrain trains a signal predictor using SVMs with pre-selected kernels for each signal. Input is a GIO and an SPF file with labeled genomic positions. The output is a TSP that can be used with SignalPredict to perform predictions on genomic sequences.

SignalPredict uses a GIO and TSP to predict signals on the given DNA sequences. The output is given in SPF.

SignalEval takes a label file and a prediction file (both SPF files) as input and computes several accuracy measures for the predictions, including the areas under the Receiver Operator Curve (ROC) and the Precision Recall Curve (PRC). This tool is useful for prediction quality monitoring.

Content prediction
Anno2ContentLabel collects labeled genomic segments for the selected content types, analogous to Anno2SignalLabel. Possible content types include 5′-UTR, exonic, intronic, 3′-UTR and intergenic. Any segment included that is not of the specified type is used as a negative example. The output is a file in CPF providing chromosome/contig name, start position, end position, strand and the label of the example.

ContentTrain is analogous to SignalTrain, with a GIO and an SPF file as inputs and a TCP object as output.

ContentPredict is analogous to SignalPredict, with a GIO and a TCP as input and an SPF file as output.

ContentEval, analogous to SignalEval, takes a CPF and SPF file as input and performs the performance evaluation.

Gene prediction
GeneTrain trains the second layer of mGene.web. Based on the GIO, genome-wide predictions for all relevant signals and content types, and a set of annotated genes, GeneTrain learns to predict gene structures from genomic DNA. The output is an internal data structure containing the TGP that can be used with GenePredict to predict genes.

GenePredict uses the TGP (either from the current history or from a list of pre-trained predictors) as well as genome-wide signal and content predictions to predict genes from the provided DNA sequences. The output is provided as a GFF3 file.

GeneEval takes two GFF3 files, one containing an annotation, the other the genome-wide gene predictions, and evaluates the prediction performance by comparing the two annotations. Note that the annotated genes should be distinct from the annotated genes used for training, otherwise the training error will be reported. Evaluation criteria include sensitivity and specificity on nucleotide, exon and gene levels ([10] and (G. Schweikert, under review) for further details).

ComposeMGenePredictor bundles all necessary trained signal, content and gene predictor objects into a TmGP that can be used with mGenePredict to predict genes.

COMPUTING TIME AND RESOURCES
The web service is running on a cluster with 84 AMD Opteron CPUs (2.2 GHz) with 8 GB of RAM per four CPUs which is shared with other web services. Training and prediction is split into several parallel sub-tasks in order to reduce the waiting time for users. Depending on the load of the cluster, whole-genome prediction of all six signals for the Caenorhabditis elegans genome (100 Mbp) takes ~24 h (about ~500 kbp/min), the prediction of the five content types also takes about 2 h (about ~1 Mbp/min), and gene prediction using these signals takes ~6 h (~300 kbp/min). The time for training the predictors strongly depends on the amount of available training data. While signal or content sensor training can typically be completed for a genome like that of C. elegans within a few hours, training the second layer of mGene.web may take 48 h (depending on the size of the training set).

For well-annotated genomes, extensive annotations exist. In these cases, we have to limit the amount of data used for training the system (due to the relatively high memory and computing time demand). In these cases,
we sub-sample the data and may therefore obtain suboptimal results compared with including all data. mGene.web currently works best for a few well-annotated regions of a genome with chromosomes/contigs of size <20 Mb, including at least a few hundred and at most a few thousand genes. Currently, the speed of the system is severely influenced by the number of chromosomes/contigs, as each one is processed separately. Moreover, there are a few built-in settings that are more suitable for organisms with compact genomes, for instance, that introns have to be shorter than 20 kb. We intend to make these options configurable in the near future.

DISCUSSION AND OUTLOOK

To the best of our knowledge, with mGene.web we currently provide the only web service that offers both prediction and complete training of a gene finding system for eukaryotic genomes. The underlying Galaxy framework greatly facilitates the data acquisition and supports workflows that appear very useful for the rather complex processes needed for genome annotations. In particular, it follows simple and standardized interfaces for data processing, training, prediction and evaluation. Advanced users may take advantage of the modular nature of mGene.web and compose their own Galaxy workflows. This should facilitate the incorporation of new developments and improvements of individual modules and will therefore be of great value for the advancement of gene finding systems.

In addition to the existing features, we will soon extend our web service in several respects. First, we will allow the utilization of additional data sources, such as EST or whole-genome alignments, as implemented in mGene.seq and mGene.multi (G. Schweikert, under review). Another line of improvements will aim at providing more advanced tools for the alignment of sequences of mRNA and orthologous proteins, thereby generating initial gene annotations for training. This will be an important step toward our next goal, to offer accurate gene prediction services for new genomes requiring no or considerably fewer known genes for training an accurate predictor.

ACKNOWLEDGEMENTS

We gratefully acknowledge the great support we received from the Galaxy development team at Penn State University, in particular James Taylor. Moreover, we would like to thank Fabio De Bona for developing an Octave interface to the MOSEK optimization toolbox as well as Peter Niermann and Zhiqin Huang for integrating new genomes and testing the system.

FUNDING

Funding for open access charge: Friedrich Miescher Laboratory of the Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES