GENE PREDICTION PROGRAM FOR PLANT GENOME

CHAN KUANG LIM; ROZANA ROSLI and
LESLIE LOW ENG TI

773

MPOB INFORMATION SERIES • ISSN 1511-7871 • JULY 2017

MPOB TS No. 173

With the advancement of next generation sequencing technologies, large numbers of sequenced genomes had been generated for many large-scale sequencing projects. Accurate prediction of genes from these genomes is one of the most important steps in the genome annotation process. Many software tools and pipelines developed by various computing techniques are available for gene prediction, but have yet to accurately predict the protein-coding regions. None of them has a universal Hidden Markov Model (HMM) that can perform gene prediction automatically for all organisms equally well. We developed an automated gene prediction program, Seqping[1] that uses self-training HMM models and transcriptomic data. The program processes the genome and transcriptome sequences of the target species using a series of gene prediction pipelines, followed by MAKER2 program to combine predictions from the tools in association with transcriptomic evidence (*Figure 1*). Seqping generates species-specific HMMs that are able to offer unbiased gene predictions. Evaluation of the program using *Oryza sativa* and *Arabidopsis thaliana* genomes showed that it was able to generate higher quality gene models (predicted genes) than those that use the standard or default HMMs in gene prediction software. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis showed that the program was able to identify at least 95% of BUSCO's plantae dataset.

THE PRODUCT

Automated gene prediction program (Seqping) that uses self-training HMM model and sequencing data to predict high quality gene models in newly sequenced plant genomes. It had been successfully used to predict large non-model plant genomes, such as the oil palm effectively. UNIX based Bash and Perl scripting were used in the program. The main script that executes a sequence of commands, including invoking other scripts written in Bash and Perl.

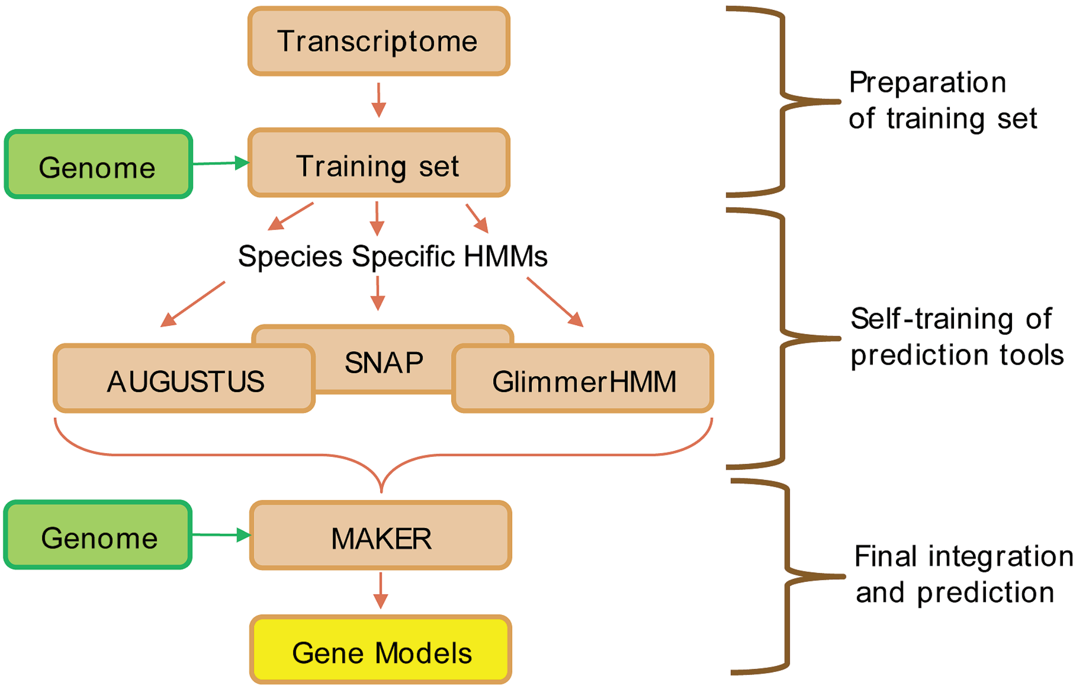

Figure 1. Seqping's gene prediction workflow.

ISSN 1511-7871
9 771511 787001

MATERIALS AND METHODS

The program consists of seven stages: (1) setting up the working directories, (2) preparation of the training set, (3) GlimmerHMM[2] training, (4) AUGUSTUS[3] training, (5) SNAP[4] training, (6) MAKER2[5] prediction, and (7) results filtering. It supports multiple processors analysis, as well as job submission to Sun Grid Engine (SGE) or Portable Batch System (PBS) job schedulers. The script's optimized parameters provide an automated and efficient tool for filtering and structural annotation of gene predictions. MAKER2, which is the final tool to combine all models (GlimmerHMM's prediction, AUGUSTUS's HMM and SNAP's HMM) and evidences (transcriptome data and NCBI Protein Database), generate the list of predicted genes in GFF format, as well as predicted gene and protein sequences in FASTA format.

NOVELTY OF THE PRODUCT

Seqping generates species-specific HMMs that are able to offer unbiased gene predictions. Our evaluation shows that Seqping was able to generate better gene predictions.

BENEFITS OF SEQPING

- Provides researchers a seamless program to train species-specific HMMs and predict genes in newly sequenced or less-studied genomes.
- Integration of multiple tools result in higher quality gene predictions in both dicotyledon and monocotyledon plants.
- The program is distributed as an open-sourced program.

ACCESSIBILITY

The first stable release of Seqping program (version: 0.1.45) has been made available to the public in June 2016 in SourceForge (https://sourceforge.net/projects/seqping/), under GNU General Public License version 3.0 (GPLv3) (*Figure 2*). It is listed as a genome annotation tool in OMICtools (https://omictools.com/seqping-tool).

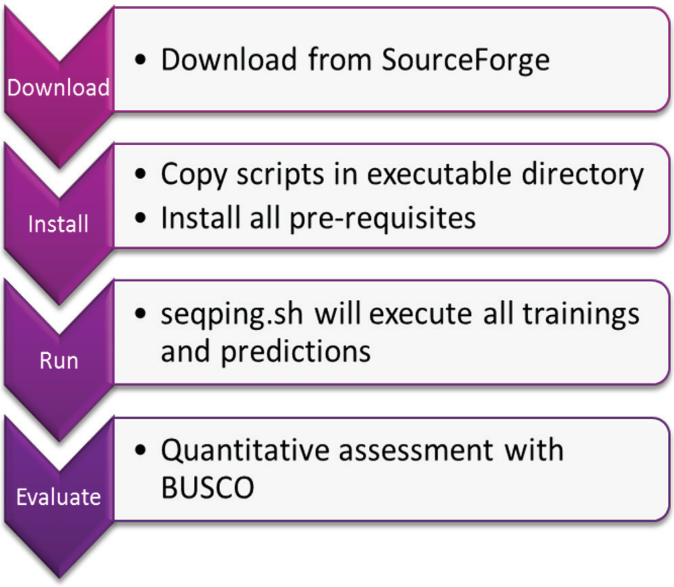
Download
• Download from SourceForge
Install
• Copy scripts in executable directory
• Install all pre-requisites
Run
• seqping.sh will execute all trainings and predictions
Evaluate
• Quantitative assessment with BUSCO

Figure 2. Basic workflow of Seqping implementation.

REFERENCES

CHAN, K-L; ROSLI, R; TATARINOVA, T; HOGAN, M; FIRDAUS-RAIH, M and LOW, E-TL (2017). Seqping: Gene prediction pipeline for plant genomes using self-trained gene models and transcriptomic data. *BMC Bioinformatics*, *18*: 29.

MAJOROS, W H H; PERTEA, M and SALZBERG, S L L (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*: 2878-2879.

STANKE, M; DIEKHANS, M; BAERTSCH, R and HAUSSLER, D (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*: 637-644.

KORF, I (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*: 59.

HOLT, C and YANDELL, M (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*: 491.

For more information, kindly contact:

Head of Corporate Implementation
and Consultancy Unit, MPOB
6, Persiaran Institusi,
Bandar Baru Bangi,
43000 Kajang, Selangor, Malaysia
Tel: 03-8769 4574
Fax: 03-8926 1337
E-mail: tot@mpob.gov.my
www.mpob.gov.my