TrainAUGUSTUS – a Web Server Application for Parameter Training and Gene Prediction in Eukaryotes

Katharina J. Hoff and Mario Stanke, University of Greifswald, GERMANY. Contact: katharina.hoff@uni-greifswald.de

Abstract

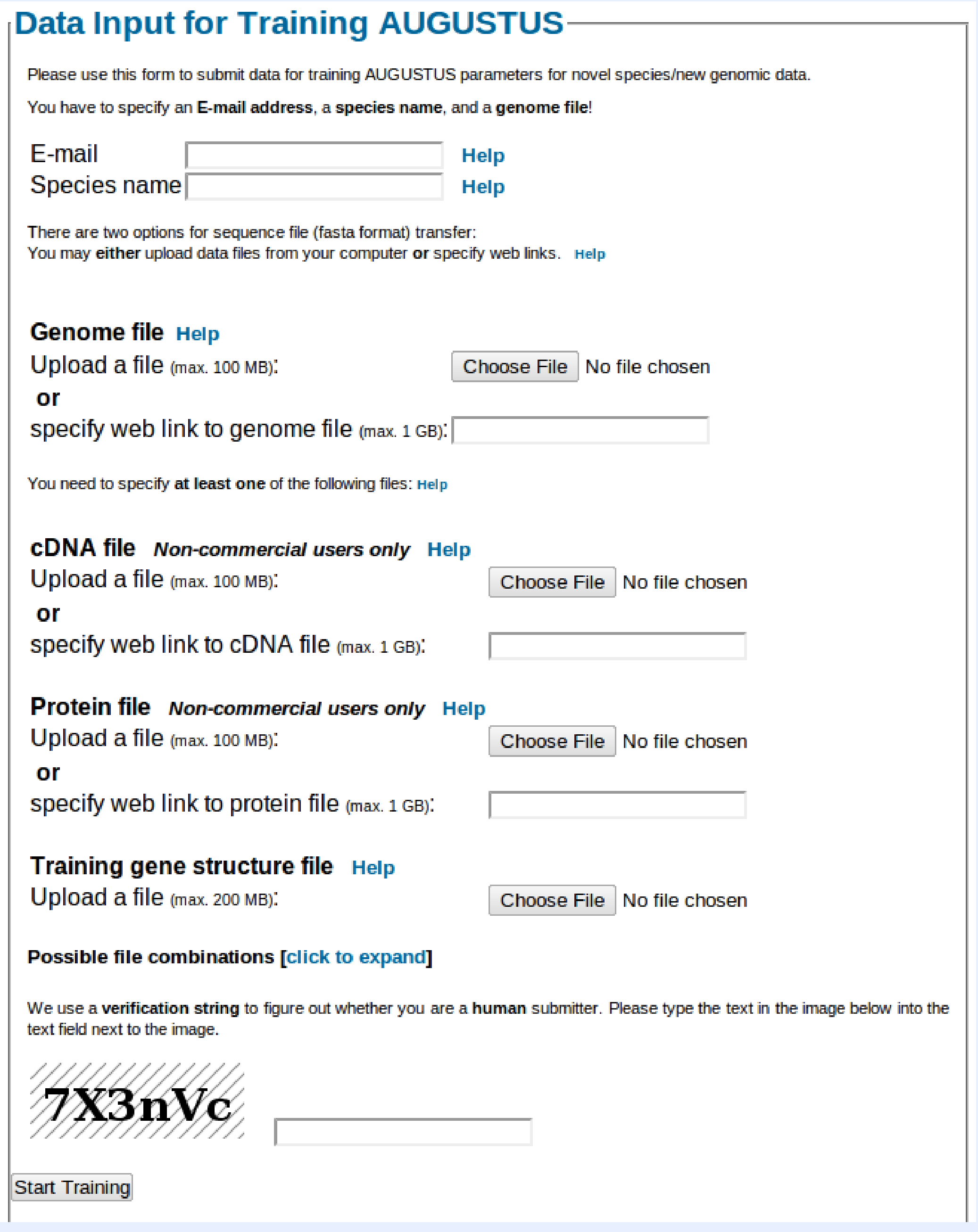
AUGUSTUS is a tool for predicting genes in eukaryotic genomic sequences [1, 2]. For achieving accurate gene predictions, a species-specific set of parameters is needed. Due to the rapidly growing number of newly sequenced genomes, an automated and easy-to-use procedure is needed in order to make gene prediction parameters for new species availabe.

Gene prediction parameters are optimized using annotated genes from the species of interest. Such initial gene sets may be generated automatically, e.g. from aligning expressed sequence tags (ESTs) to genomic sequences, or by mapping protein coding genes from other species to the genome.

We present a web server application for creating high quality training gene sets from ESTs or protein sequences. Subsequent to finding training genes, the web server application optimizes AUGUSTUS parameters and makes predictions in the supplied genomic sequence using the newly trained parameters and the supplied ESTs or protein sequences as external supporting evidence ("hints"). It is also possible to supply hints that were created externally, e.g. through manual editing, or from RNAseq data alignments.

The web server application is available at http://bioinf.uni-greifswald.de/trainaugustus

AUGUSTUS Training Web Interface



Data Input for Training AUGUSTUS

Please use this form to submit data for training AUGUSTUS parameters for novel species/new genomic data.

You have to specify an **E-mail address**, a **species name**, and a **genome file**!

E-mail Help
Species name Help

There are two options for sequence file (fasta format) transfer:
You may **either** upload data files from your computer **or** specify web links. Help

Genome file Help
Upload a file (max. 100 MB): Choose File No file chosen
or
specify web link to genome file (max. 1 GB):

You need to specify **at least one** of the following files: Help

cDNA file ***Non-commercial users only*** Help
Upload a file (max. 100 MB): Choose File No file chosen
or
specify web link to cDNA file (max. 1 GB):

Protein file ***Non-commercial users only*** Help
Upload a file (max. 100 MB): Choose File No file chosen
or
specify web link to protein file (max. 1 GB):

Training gene structure file Help
Upload a file (max. 200 MB): Choose File No file chosen

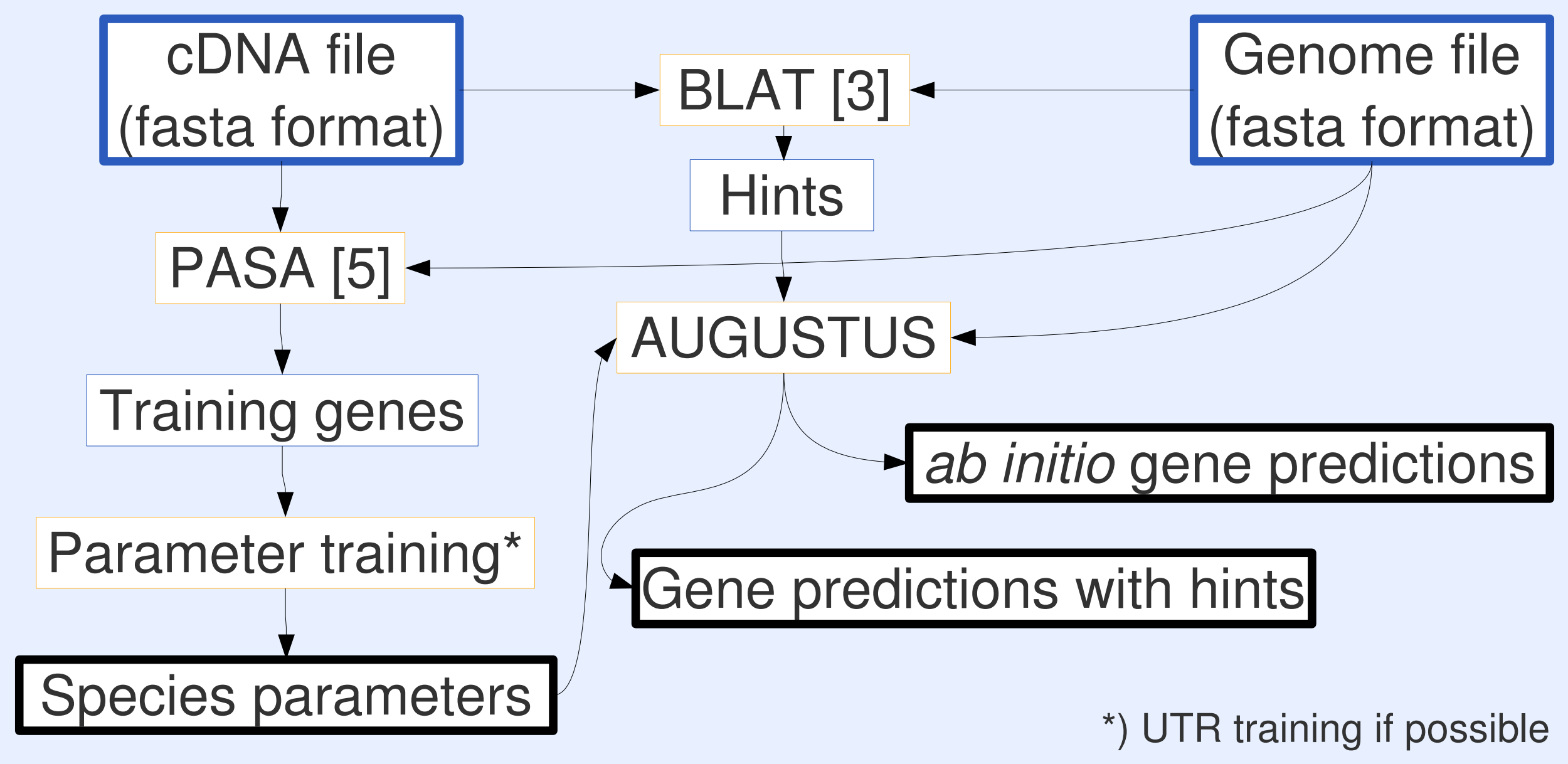
Possible file combinations [click to expand]

We use a **verification string** to figure out whether you are a **human** submitter. Please type the text in the image below into the text field next to the image.

7X3nVc

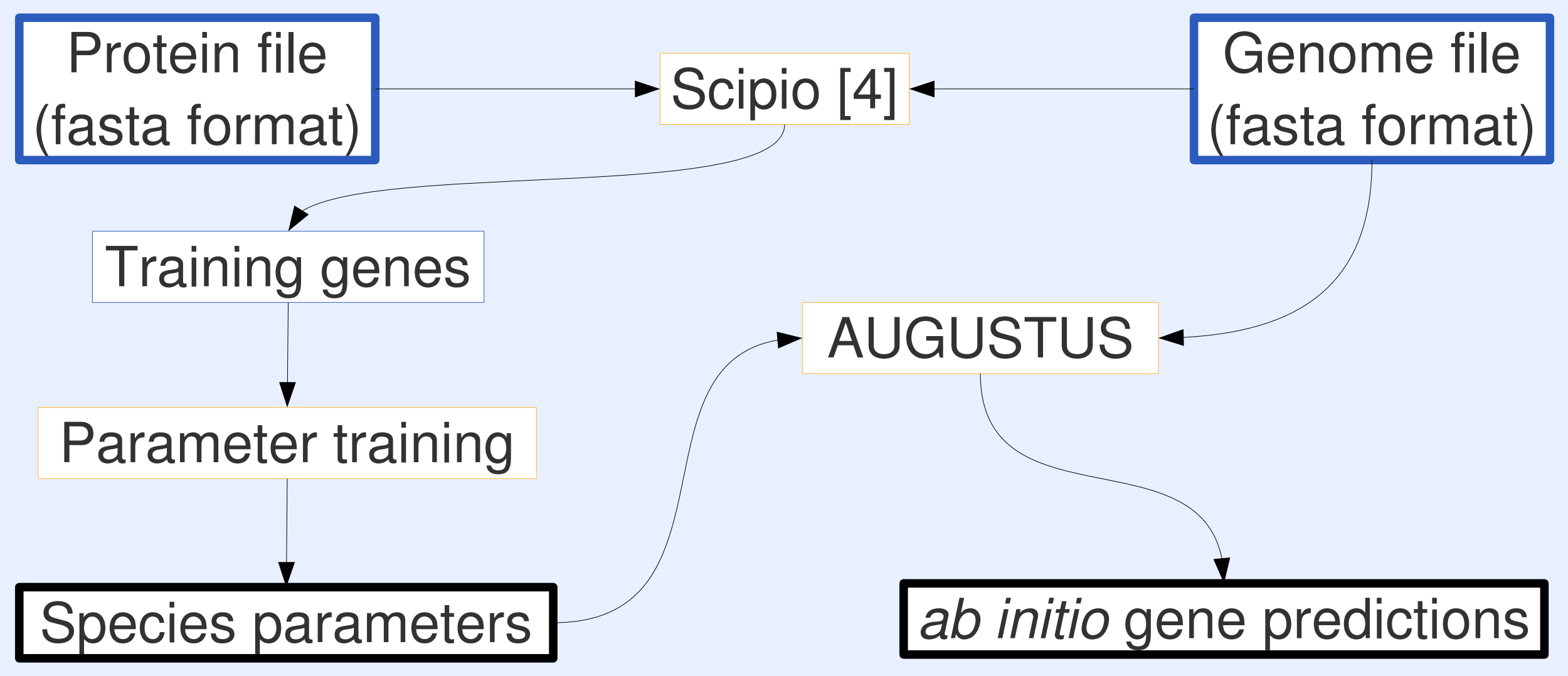
Start Training

Training: cDNA and Genome File

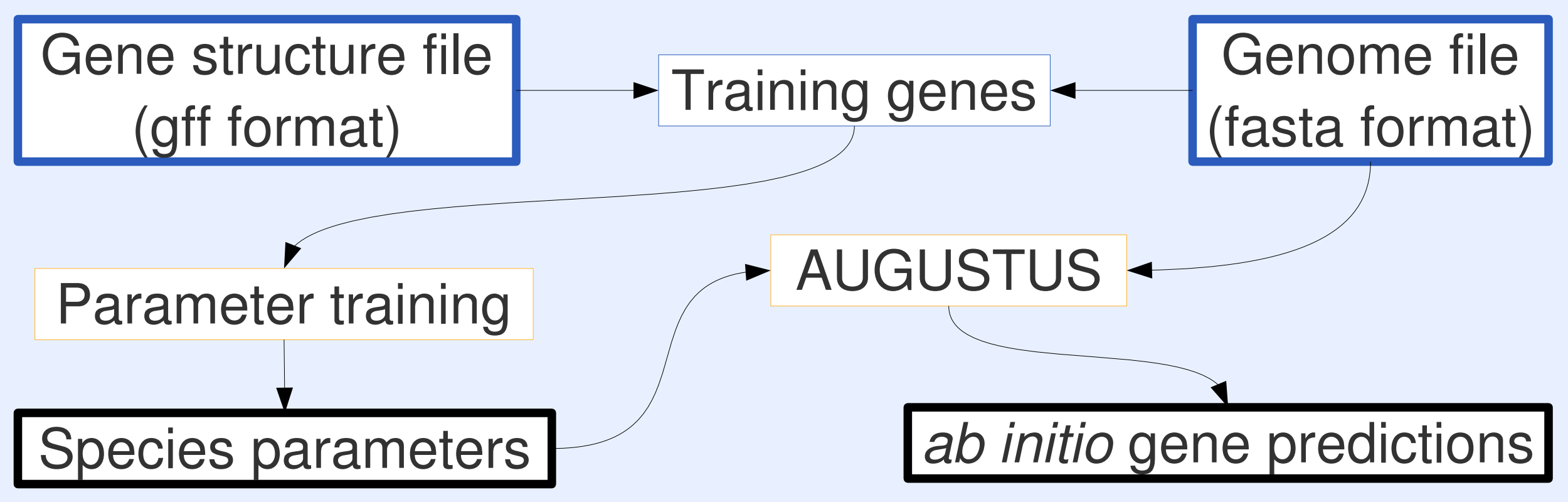


*) UTR training if possible

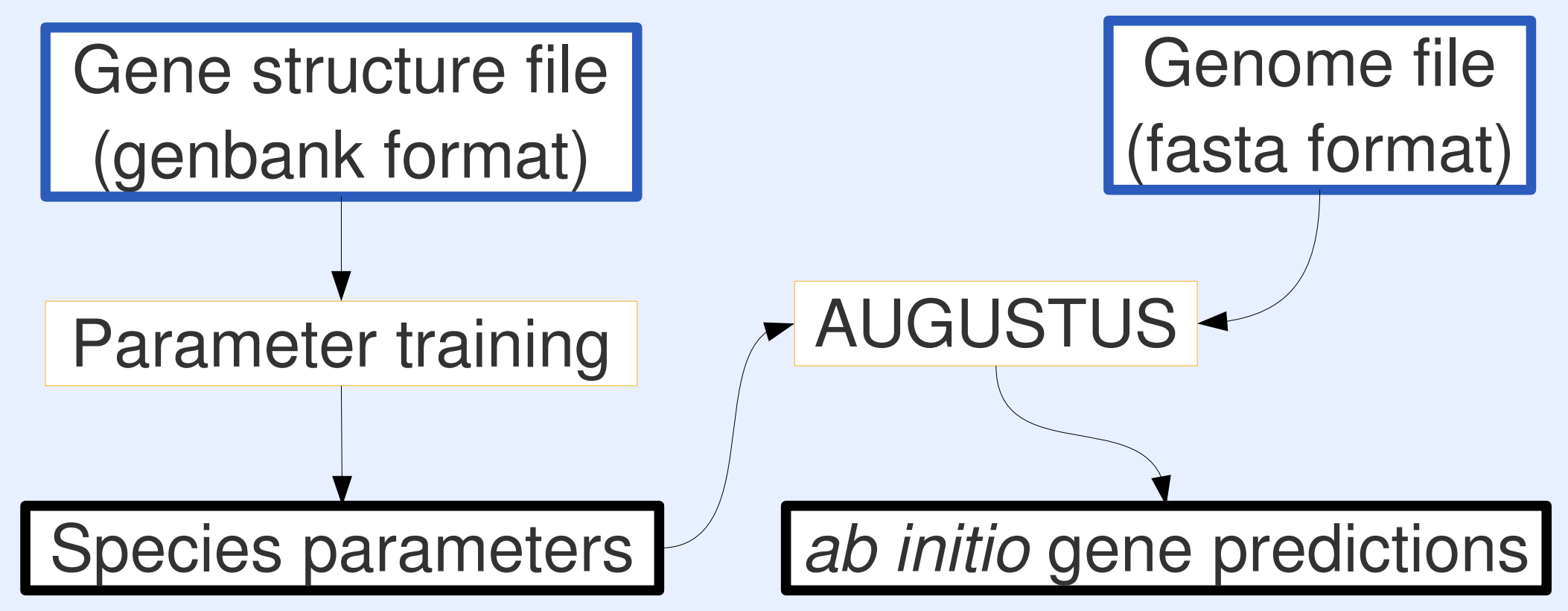
Training: Protein and Genome File


Protein file (fasta format)
Scipio [4]
Genome file (fasta format)
Training genes
AUGUSTUS
Parameter training
Species parameters
ab initio gene predictions

Training: Gene Structure and Genome File (gff option)


Gene structure file (gff format)
Training genes
Genome file (fasta format)
Parameter training
AUGUSTUS
Species parameters
ab initio gene predictions

Training: Gene Structure and Genome File (gb option)


Gene structure file (genbank format)
Genome file (fasta format)
Parameter training
AUGUSTUS
Species parameters
ab initio gene predictions

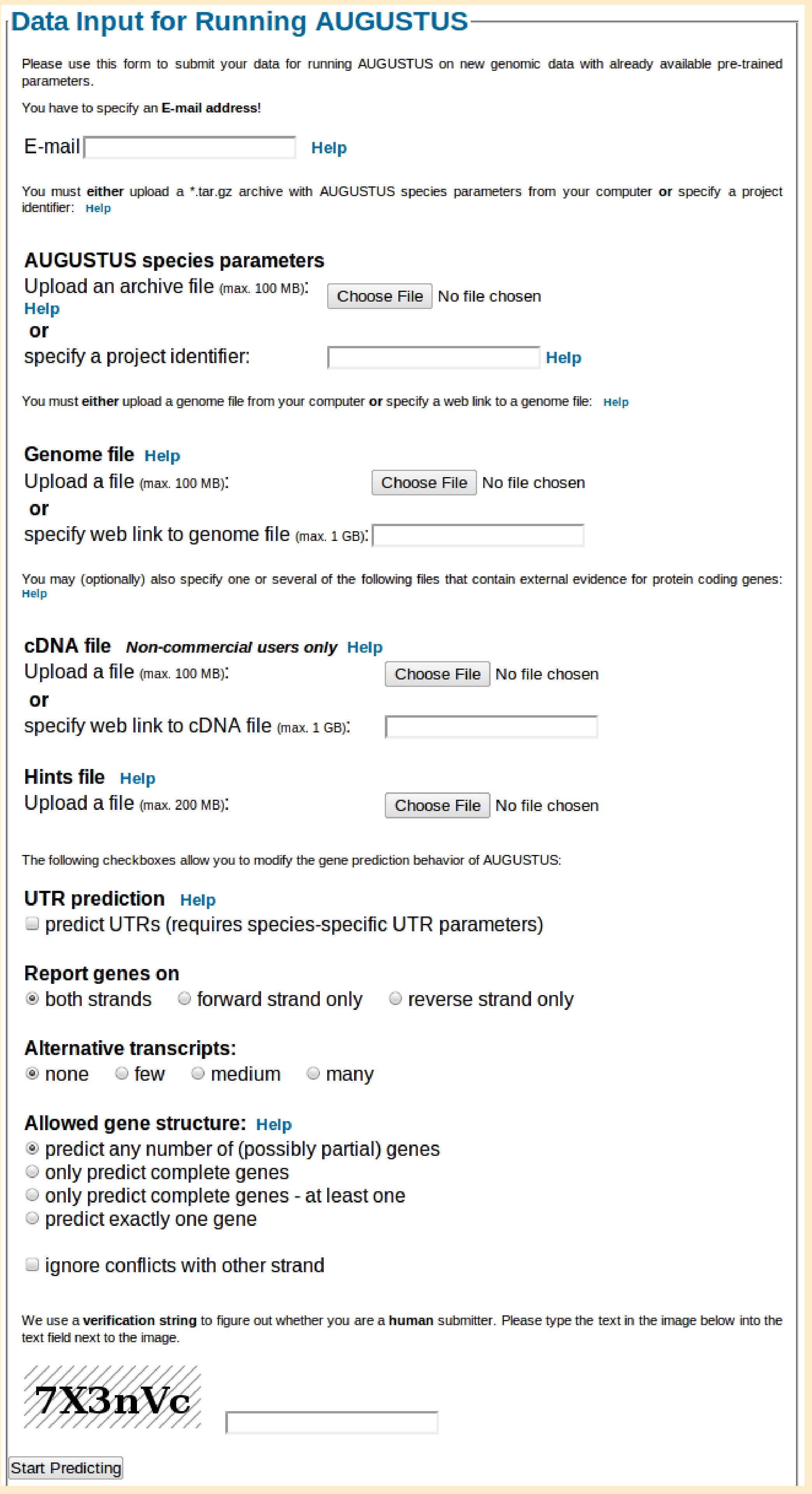
Training: cDNA, Protein and Genome File

Training gene generation see “Protein and Genome File”. **Hints file generation** see “cDNA and Genome File”.
Results: *ab initio* gene predictions and gene predictions with hints.

Training: cDNA, Gene Structure and Genome File

Training gene generation see “Gene Structure and Genome File”. **Hints file generation** see “cDNA and Genome File”.
Results: *ab initio* gene predictions and gene predictions with hints.

AUGUSTUS Prediction Web Interface



Data Input for Running AUGUSTUS

Please use this form to submit your data for running AUGUSTUS on new genomic data with already available pre-trained parameters.

You have to specify an **E-mail address!**

E-mail Help

You must **either** upload a *.tar.gz archive with AUGUSTUS species parameters from your computer **or** specify a project identifier: Help

AUGUSTUS species parameters
Upload an archive file (max. 100 MB): Choose File No file chosen
Help
or
specify a project identifier: Help

You must **either** upload a genome file from your computer **or** specify a web link to a genome file: Help

Genome file Help
Upload a file (max. 100 MB): Choose File No file chosen
or
specify web link to genome file (max. 1 GB):

You may (optionally) also specify one or several of the following files that contain external evidence for protein coding genes: Help

cDNA file ***Non-commercial users only*** Help
Upload a file (max. 100 MB): Choose File No file chosen
or
specify web link to cDNA file (max. 1 GB):

Hints file Help
Upload a file (max. 200 MB): Choose File No file chosen

The following checkboxes allow you to modify the gene prediction behavior of AUGUSTUS:

UTR prediction Help
□ predict UTRs (requires species-specific UTR parameters)

Report genes on
● both strands ○ forward strand only ○ reverse strand only

Alternative transcripts:
● none ○ few ○ medium ○ many

Allowed gene structure: Help
● predict any number of (possibly partial) genes
○ only predict complete genes
○ only predict complete genes - at least one
○ predict exactly one gene

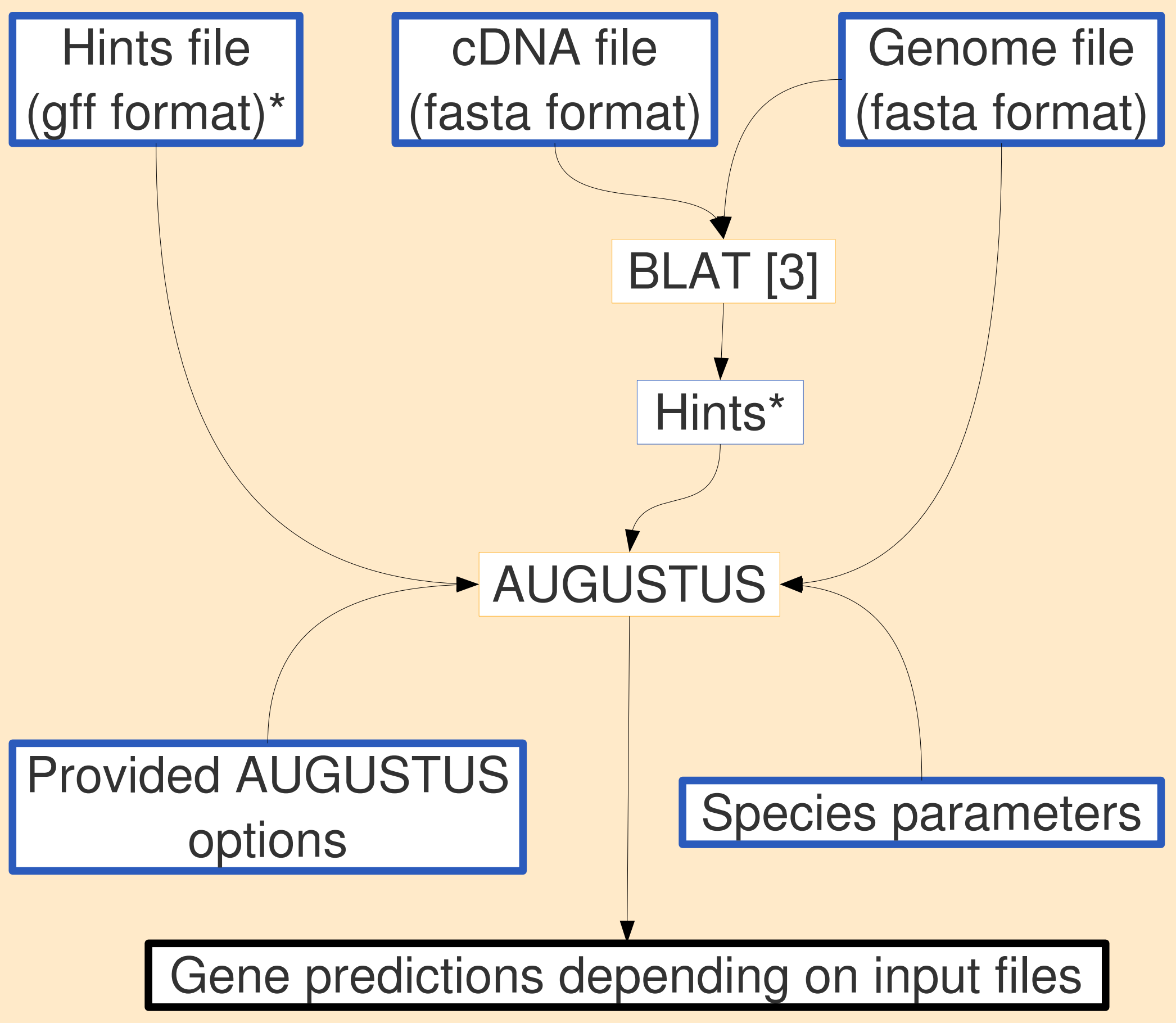
□ ignore conflicts with other strand

We use a **verification string** to figure out whether you are a **human** submitter. Please type the text in the image below into the text field next to the image.

7X3nVc

Start Predicting

Prediction Pipeline


Hints file (gff format)*
cDNA file (fasta format)
Genome file (fasta format)
BLAT [3]
Hints*
AUGUSTUS
Provided AUGUSTUS options
Species parameters
Gene predictions depending on input files

A project identifier is assigned to each AUGUSTUS training run. This ID may be used to call pre-trained parameters. The upload of external parameters is also possible.

*) Hints can be provided in a file, and hints are generated from provided cDNA sequence data by the prediction pipeline. Externally provided hints are treated by AUGUSTUS as “manually created”, i.e. they have a stronger influence on predictions than hints that are generated by the web server application pipeline.

Results of TrainAUGUSTUS

Each training run produces at least the following files:
- AutoAug.log → an event log file
- AutoAug.err → an error log file

In addition, the following files may be produced depending on the input file combination and depending on training success:
- parameters.tar.gz → archive with AUGUSTUS species parameters
- training.gb.gz → gzipped genbank file with produced training genes
- ab_initio.tar.gz → archive with ab initio gene predictions
- hints_pred.tar.gz → archive that contains gene predictions with hints

Gene prediction archives contain at least a file that contains gene predictions in gff format.
Additionally, a gtf-file and fasta files with amino acid sequences, exon sequences, coding sequences and mRNA sequences may be included.
Also a gbrowse-file may be produced.

References

[1] M. Stanke and S. Waack (2003) "Gene prediction with a hidden Markov model and a new intron submodel", Bioinformatics, Vol. 19, Suppl. 2, pages ii215-ii225
[2] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler (2008) "Using native and syntenically mapped cDNA alignments to improve de novo gene finding", Bioinformatics, 24(5):637
[3] Kent, W.J. (2002) BLAT—The BLAST-Like Alignment Tool. Genome Res, 12, 656-664.
[4] Keller, O., Odronitz, F., Stanke, M., Kollmar, M., Waack, S. (2008) Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9, 278.
[5] Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res, 31, 5654-5666.

Funding

We thank the German Research Foundation (DFG) for funding the development of this web application through the grants STA 1009/4 and HO 4545/1-1.