

## §S10. Gene Prediction Protocol and Accuracy Estimation

### §S10.1. Gene Prediction Protocol

Gene structures were predicted using a combination of FGENESH, FGENESH+, and GENEWISE. Both FGENESH and FGENESH+ are gene prediction programs acquired from Softberry.com and GENEWISE is part of the WISE2 package developed by Ewan Birney and is available from the Sanger Center.

Both FGENESH and FGENESH+ utilize a statistical model of gene structure that require training on each organism for accurate prediction. FGENESH+ additionally combines a protein sequence with the statistical model to improve accuracy. We acquired these programs already trained by Softberry on *Aspergillus nidulans* sequences.

GENEWISE (as we ran it), splices and aligns a protein sequence with genomic sequence to predict a gene structure. Although GENEWISE does utilize some species-specific parameters, most notably for intron nucleotide statistics and splice site consensus sequences, these can be set to non-species specific defaults. In this case, GENEWISE essentially produces the best local alignment of a protein assuming that introns start at GT and end at AG most of the time and in some cases this results a full alignment of the protein to the genome. Since we are interested in predicting complete gene structures, we post-processed GENEWISE incomplete protein alignments by moving the first and last exon upstream or downstream to the nearest start and stop codons respectively. If a stop codon was encountered upstream of a gene before a start could be found, the gene call was not used.

Briefly, these three gene callers were combined in the following manner:

1. FGENESH was run on the entire genomic sequence to provide an initial set of predicted genes. Each FGENESH predicted was put into a set of EVIDENCE\_GENES.
2. The genome was also searched against the non-redundant protein database using BLASTX
3. Regions of the genome with blastx homology spanning over 80% of a protein (when sub-alignments are stitched together in a consistent fashion) were considered "Homologous Gene Regions" (HGRs).
4. HGRs were clustered into groups of HGRs that all implicated the same gene structure (most often representing groups of essentially orthologous proteins).
5. For each cluster of HGRs, the protein showing the most sequence similarity to the genome was passed to both FGENESH and GENEWISE to produce 2 gene predictions, if the protein had >80% amino acid identity to the translated genome (cumulative across sub-alignments).
6. If the protein used in the previous had >90% amino acid identity to the translated genome (cumulative across sub-alignments), then the GENEWISE call, if valid, was favored over the FGENESH+ call, and was used as the EVIDENCE\_GENE for the HGR (see below for the reason why) and added to the set of EVIDENCE\_GENES. If this protein had >80% but less than 90% amino acid identity to the translated genome (cumulative across

sub-alignments), then the FGENESH+ call, if valid, was favored over the GENEWISE call, and was used as the EVIDENCE\_GENE for the HGR and added to the set of EVIDENCE\_GENES.

7. When EVIDENCE\_GENES overlapped in their exons, the EVIDENCE\_GENE with the least amount of homology support (as measured by the sequence similarity of the protein used to make the call or zero for FGENESH calls) was removed from the set of EVIDENCE\_GENES.
8. All remaining EVIDENCE\_GENES were then called as our official ANNOTATED\_GENES and passed to the next step of gene calling for Gene Naming.

Additional information is available

[http://www.broad.mit.edu/annotation/fungi/aspergillus/gene\\_finding.html](http://www.broad.mit.edu/annotation/fungi/aspergillus/gene_finding.html)

## §S10.2. Accuracy Estimation

The accuracy of our gene prediction protocol was assessed by comparing predicted to overlapping EST alignments (aligned using est2genome). Alignments of individual ESTs were grouped into *EST clusters* corresponding to sets of ESTs predicting the identical gene structure at a particular locus.

**Table §S10.1 – Estimated Gene Prediction Accuracy**

Category	Number	Percentage
Genes	9541	
EST Clusters		
EST Clusters Not Overlapping Genes	946	20%
Genes Overlapping EST Clusters	3143	33%
Genes w/No Discrepancies Relative to EST Alignments	2183	69%
Genes hitting multiple EST clusters	563	
Genes w/Discrepancies		
Genes w/Missing Exons	119	4%
Genes w/Wrong Exons	35	1%
Genes w/Splice Junction Difference	907	29%
Nucleotide level statistics relative to EST alignments (Bursset and Guigo 1996)		
AC (Approximate Correlation)	0.76	
ACP (Average Conditional Probability)	0.88	
CC (Correlation Coefficient)	0.76	
Sensitivity	0.94	
Specificity	0.98	