

Eukaryotic GeneMark.hmm ES-3.0 and E-3.0 Accuracy Comparison

Gene structure prediction*

Table 1. Values of several categories of sensitivity and specificity (Sn/Sp) and (Sn+Sp)/2 characterizing the accuracy of gene predictions produced for the group of 'well-studied' genomes by the eukaryotic GeneMark.hmm with models derived by both unsupervised and supervised training

	<i>A. thaliana</i>				<i>C. elegans</i>				<i>D. melanogaster</i>			
	unsupervised		supervised		unsupervised		supervised		unsupervised		supervised	
Nucleotide	97.7 94.8	96.3	97.2 94.3	95.8	99.1 93.6	96.4	97.8 95.5	96.7	97.9 92.9	95.4	98.1 93.1	95.6
Internal exons	91.2 87.8	89.5	91.2 88.5	89.9	94.0 91.3	92.7	90.9 90.8	90.9	91.3 89.7	90.5	87.2 90.2	88.7
Initiation sites	80.1 76.5	78.3	80.1 71.9	76.0	85.8 68.9	77.4	79.2 67.4	73.3	83.9 73.5	78.7	83.4 74.3	78.9
Termination sites	87.5 83.1	85.3	88.3 78.6	83.5	95.1 75.3	85.2	94.0 79.6	86.8	89.2 77.2	83.2	89.5 78.8	84.2
Donor sites	94.0 90.3	92.2	94.0 89.8	91.9	96.2 90.8	93.5	93.7 91.4	92.6	92.8 87.2	90.0	91.3 89.1	90.2
Acceptor sites	94.0 90.2	92.1	93.6 89.2	91.4	97.3 91.6	94.5	95.2 92.8	94.0	93.0 87.0	90.0	90.5 87.9	89.2

Boldface highlights the higher value in comparison of unsupervised and supervised modes (ES-3.0 versus E-3.0).

Table 2. Same as in Table 1, for the group of novel genomes and the unsupervised mode only (GeneMark.hmm ES-3.0)

	<i>A. gambiae</i>	<i>C. intestinalis</i>	<i>C. reinhardtii</i>	<i>T. gondii</i>				
Nucleotides	96.0 85.0	90.5	98.3 90.0	94.2	97.4 97.4	97.4	89.6 87.1	88.4
Internal exons	89.3 88.4	88.9	94.8 92.1	93.5	91.4 95.4	93.4	80.2 83.1	81.7
Initiation sites	77.8 67.9	72.9	79.6 63.0	71.3	82.9 73.9	78.4	58.5 71.7	65.1
Termination sites	86.1 71.7	78.9	85.4 66.3	75.9	92.7 82.6	87.7	66.2 81.1	73.7
Donor sites	89.7 84.1	86.9	95.3 89.7	92.5	94.1 96.3	95.2	81.3 87.5	84.4
Acceptor sites	92.3 84.7	88.5	96.3 90.3	93.3	93.5 95.7	94.6	82.0 88.3	85.2

Table 3. Values of nucleotide level sensitivity and specificity (Sn/Sp) along with (Sn + Sp)/2 for gene predictions produced by GeneMark.hmm and SNAP for the group of 'well-known' genomes.

	GeneMark.hmm unsupervised (ES-3.0)		SNAP bootstrap		GeneMark.hmm supervised (E-3.0)		SNAP supervised		GenScan supervised		Genfinder supervised		Augustus supervised	
<i>A. thaliana</i>	98.3 94.7	96.5	96.6 93.2	94.9	98.4 94.2	96.3	97.1 95.2	96.2	79.9 92.9	86.4	- -	- -	- -	- -
<i>C. elegans</i>	99.1 95.1	97.1	96.7 91.1	93.9	97.7 96.2	97.0	97.6 94.2	95.9	- -	-	98.1 95.3	96.7	- -	- -
<i>D. melanogaster</i>	93.8 86.1	90.0	92.5 86.6	89.6	93.2 87.7	90.5	94.3 86.5	90.4	- -	-	- -	- -	92.4 88.6	90.5

The unsupervised mode of GeneMark.hmm(ES-3.0) is compared with the bootstrapped mode of SNAP (first two columns). The performance of the supervised modes of the two programs, GeneMark.hmmE-3 and SNAP, are also compared (last five columns) with performance of GenScan (trained on *A.thaliana*), Genfinder (trained on *C.elegans*) and Augustus (trained on *D.melanogaster*). The accuracy of GeneMark.hmm was assessed on the same test sets downloaded from the SNAP website.

Table 4. Values of sensitivity and specificity (Sn/Sp) along with (Sn + Sp)/2 for gene predictions produced by GeneMark.hmm and SNAP for the test sets described in the Dataset section

	<i>A. thaliana</i>				<i>A. gambiae</i>				<i>C. intestinalis</i>				<i>C. elegans</i>				<i>D. melanogaster</i>			
	GeneMark.hmm unsupervised		SNAP supervised		GeneMark.hmm unsupervised		SNAP supervised		GeneMark.hmm unsupervised		SNAP supervised		GeneMark.hmm unsupervised		SNAP supervised		GeneMark.hmm unsupervised		SNAP supervised	
Nucleotide	<u>97.7</u> 94.8	96.3	<u>93.6</u> 95.3	94.5	<u>96.0</u> 85.0	90.5	<u>87.6</u> 81.4	84.5	<u>98.3</u> 90.0	94.2	<u>90.1</u> 74.3	82.2	<u>99.1</u> 93.6	96.4	<u>97.2</u> 94.1	95.7	<u>97.9</u> 92.9	95.4	<u>94.8</u> 92.9	93.9
Internal exons	<u>91.2</u> 87.8	89.5	<u>79.7</u> 87.8	83.8	<u>89.3</u> 88.4	88.9	<u>81.7</u> 87.6	84.7	<u>94.8</u> 92.1	93.5	<u>80.9</u> 61.7	71.3	<u>94.0</u> 91.3	92.7	<u>87.1</u> 87.1		<u>91.3</u> 89.7	90.5	<u>85.8</u> 85.2	85.5
Initiation sites	<u>80.1</u> 76.5	78.3	<u>75.6</u> 74.3	75.0	<u>77.8</u> 67.9	72.9	<u>65.2</u> 67.2	66.2	<u>79.6</u> 63.0	71.3	<u>61.0</u> 43.4	52.2	<u>85.8</u> 68.9	77.4	<u>73.2</u> 61.5	67.4	<u>83.9</u> 73.5	78.7	<u>78.1</u> 77.3	77.7
Termination sites	<u>87.5</u> 83.1	85.3	<u>84.0</u> 82.9	83.5	<u>86.1</u> 71.7	78.9	<u>78.8</u> 73.2	76.0	<u>85.4</u> 66.3	75.9	<u>63.3</u> 45.9	54.6	<u>95.1</u> 75.3	85.2	<u>89.4</u> 72.4	80.9	<u>89.2</u> 77.2	83.2	<u>78.1</u> 76.8	77.5
Donor sites	<u>94.0</u> 90.3	92.2	<u>83.7</u> 90.0	86.9	<u>89.7</u> 84.1	86.9	<u>80.8</u> 85.3	83.1	<u>95.3</u> 89.7	92.5	<u>82.7</u> 62.4	72.6	<u>96.2</u> 90.8	93.5	<u>90.1</u> 87.8	89.0	<u>92.8</u> 87.2	90.0	<u>86.9</u> 86.2	86.6
Acceptor sites	<u>94.0</u> 90.2	92.1	<u>84.6</u> 91.0	87.8	<u>92.3</u> 84.7	88.5	<u>83.3</u> 84.1	83.7	<u>96.3</u> 90.3	93.3	<u>83.6</u> 63.3	73.5	<u>97.3</u> 91.6	94.5	<u>93.5</u> 90.6	92.1	<u>93.0</u> 87.0	90.0	<u>87.7</u> 86.8	87.3

The unsupervised mode of Gene-Mark.hmm (ES-3.0) is compared with the supervised mode of SNAP which typically performs better than the SNAP bootstrapped mode (6). The supervised models from SNAP were downloaded from the SNAP website.

*Lomsadze A., Ter-Hovhannisyan V., Chernoff Y. and Borodovsky M.
 "Gene identification in novel eukaryotic genomes by self-training algorithm"
Nucleic Acids Research, 2005, Vol. 33, No. 20, 6494-6506