ID: 1
Name: Drosophila
NCBI Sequence: AH005351.3

==========AUGUSTUS==========

----- prediction on sequence number 1 (length = 3646, name = unnamed-1) -----
#
Constraints/Hints:
(none)
Predicted genes for sequence number 1 on both strands
gene g1
AUGUSTUS gene 56 3547 0.07 +

```
# start gene g1
unnamed-1       AUGUSTUS        gene    56      3547    0.07    +       .       g1
unnamed-1       AUGUSTUS        transcript      56      3547    0.07    +       .       g1.t1
unnamed-1       AUGUSTUS        tss     56      56      .       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        exon    56      332     .       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        start_codon     89      91      .       +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        initial 89      332     0.77    +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        internal        647     2394    0.98    +       2       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        terminal        2967    3218    1       +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        intron  333     646     1       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        intron  2395    2966    1       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        CDS     89      332     0.77    +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        CDS     647     2394    0.98    +       2       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        exon    647     2394    .       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        CDS     2967    3218    1       +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        exon    2967    3547    .       +       .       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        stop_codon      3216    3218    .       +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        tts     3547    3547    .       +       .       transcript_id "g1.t1"; gene_id "g1";
```

==========GENEMARK==========

Predicted genes/exons

Gene #	Exon #	Strand	Exon Type	Exon Range		Exon Length	Start/End Frame
1	1	+	Initial	89	332	244	1 1 - -
1	2	+	Internal	647	2394	1748	2 3 - -
1	3	+	Terminal	2967	3218	252	1 3 - -

ID: 2
Name: E. coli
NCBI Sequence: NZ_CCRE01000039.1

==========AUGUSTUS==========

----- prediction on sequence number 1 (length = 3761, name = unnamed-1) -----
#
Constraints/Hints:

(none)
Predicted genes for sequence number 1 on both strands
gene g1
AUGUSTUS gene 96 599 0.49 +
gene g2
AUGUSTUS gene 707 2065 0.45 -
gene g3
AUGUSTUS gene 2065 2847 0.65 -
gene g4
AUGUSTUS gene 2869 3282 1 +
gene g5
AUGUSTUS gene 3391 3660 0.97 +

```
# start gene g1
unnamed-1       AUGUSTUS        gene    96      599     0.49    +       .       g1
unnamed-1       AUGUSTUS        transcript      96      599     0.49    +       .       g1.t1
unnamed-1       AUGUSTUS        start_codon     96      98      .       +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        single  96      599     0.49    +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        CDS     96      599     0.49    +       0       transcript_id "g1.t1"; gene_id "g1";
unnamed-1       AUGUSTUS        stop_codon      597     599     .       +       0       transcript_id "g1.t1"; gene_id "g1";
# start gene g2
unnamed-1       AUGUSTUS        gene    707     2065    0.45    -       .       g2
unnamed-1       AUGUSTUS        transcript      707     2065    0.45    -       .       g2.t1
unnamed-1       AUGUSTUS        stop_codon      707     709     .       -       0       transcript_id "g2.t1"; gene_id "g2";
unnamed-1       AUGUSTUS        single  707     2065    0.45    -       0       transcript_id "g2.t1"; gene_id "g2";
unnamed-1       AUGUSTUS        CDS     707     2065    0.45    -       0       transcript_id "g2.t1"; gene_id "g2";
unnamed-1       AUGUSTUS        start_codon     2063    2065    .       -       0       transcript_id "g2.t1"; gene_id "g2";
# start gene g3
unnamed-1       AUGUSTUS        gene    2065    2847    0.65    -       .       g3
unnamed-1       AUGUSTUS        transcript      2065    2847    0.65    -       .       g3.t1
unnamed-1       AUGUSTUS        stop_codon      2065    2067    .       -       0       transcript_id "g3.t1"; gene_id "g3";
unnamed-1       AUGUSTUS        single  2065    2847    0.65    -       0       transcript_id "g3.t1"; gene_id "g3";
unnamed-1       AUGUSTUS        CDS     2065    2847    0.65    -       0       transcript_id "g3.t1"; gene_id "g3";
unnamed-1       AUGUSTUS        start_codon     2845    2847    .       -       0       transcript_id "g3.t1"; gene_id "g3";
# start gene g4
unnamed-1       AUGUSTUS        gene    2869    3282    1       +       .       g4
unnamed-1       AUGUSTUS        transcript      2869    3282    1       +       .       g4.t1
unnamed-1       AUGUSTUS        start_codon     2869    2871    .       +       0       transcript_id "g4.t1"; gene_id "g4";
unnamed-1       AUGUSTUS        single  2869    3282    1       +       0       transcript_id "g4.t1"; gene_id "g4";
unnamed-1       AUGUSTUS        CDS     2869    3282    1       +       0       transcript_id "g4.t1"; gene_id "g4";
unnamed-1       AUGUSTUS        stop_codon      3280    3282    .       +       0       transcript_id "g4.t1"; gene_id "g4";
# start gene g5
unnamed-1       AUGUSTUS        gene    3391    3660    0.97    +       .       g5
unnamed-1       AUGUSTUS        transcript      3391    3660    0.97    +       .       g5.t1
unnamed-1       AUGUSTUS        start_codon     3391    3393    .       +       0       transcript_id "g5.t1"; gene_id "g5";
unnamed-1       AUGUSTUS        single  3391    3660    0.97    +       0       transcript_id "g5.t1"; gene_id "g5";
unnamed-1       AUGUSTUS        CDS     3391    3660    0.97    +       0       transcript_id "g5.t1"; gene_id "g5";
unnamed-1       AUGUSTUS        stop_codon      3658    3660    .       +       0       transcript_id "g5.t1"; gene_id "g5";
```

==========GENEMARK==========

Model information: Escherichia_coli_K_12_substr__MG1655

FASTA definition line: empty-fasta-def-line
Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	72	599	528	1
2	-	707	2065	1359	2
3	-	2065	2784	720	2
4	+	2869	3282	414	2
5	+	3391	3660	270	1

ID: 3
Name: Staphylococcus
NCBI Sequence: L19300.1

==========AUGUSTUS==========

----- prediction on sequence number 1 (length = 3646, name = unnamed-1) -----
#
Constraints/Hints:
(none)
Predicted genes for sequence number 1 on both strands
gene g1
AUGUSTUS gene 462 572 0.94 +
gene g2
AUGUSTUS gene 574 654 0.91 +
gene g3
AUGUSTUS gene 1798 2313 1 +
gene g4
AUGUSTUS gene 2651 3100 1 +
gene g5
AUGUSTUS gene 3066 3941 0.68 -

```
# start gene g1
unnamed-1	AUGUSTUS	gene	462	572	0.94	+	.	g1
unnamed-1	AUGUSTUS	transcript	462	572	0.94	+	.	g1.t1
unnamed-1	AUGUSTUS	start_codon	462	464	.	+	0	transcript_id "g1.t1"; gene_id "g1";
unnamed-1	AUGUSTUS	single	462	572	0.94	+	0	transcript_id "g1.t1"; gene_id "g1";
unnamed-1	AUGUSTUS	CDS	462	572	0.94	+	0	transcript_id "g1.t1"; gene_id "g1";
unnamed-1	AUGUSTUS	stop_codon	570	572	.	+	0	transcript_id "g1.t1"; gene_id "g1";
# start gene g2
unnamed-1	AUGUSTUS	gene	574	654	0.91	+	.	g2
unnamed-1	AUGUSTUS	transcript	574	654	0.91	+	.	g2.t1
unnamed-1	AUGUSTUS	start_codon	574	576	.	+	0	transcript_id "g2.t1"; gene_id "g2";
unnamed-1	AUGUSTUS	single	574	654	0.91	+	0	transcript_id "g2.t1"; gene_id "g2";
unnamed-1	AUGUSTUS	CDS	574	654	0.91	+	0	transcript_id "g2.t1"; gene_id "g2";
unnamed-1	AUGUSTUS	stop_codon	652	654	.	+	0	transcript_id "g2.t1"; gene_id "g2";
# start gene g3
unnamed-1	AUGUSTUS	gene	1798	2313	1	+	.	g3
unnamed-1	AUGUSTUS	transcript	1798	2313	1	+	.	g3.t1
unnamed-1	AUGUSTUS	start_codon	1798	1800	.	+	0	transcript_id "g3.t1"; gene_id "g3";
unnamed-1	AUGUSTUS	single	1798	2313	1	+	0	transcript_id "g3.t1"; gene_id "g3";
unnamed-1	AUGUSTUS	CDS	1798	2313	1	+	0	transcript_id "g3.t1"; gene_id "g3";
unnamed-1	AUGUSTUS	stop_codon	2311	2313	.	+	0	transcript_id "g3.t1"; gene_id "g3";
# start gene g4
unnamed-1	AUGUSTUS	gene	2651	3100	1	+	.	g4
unnamed-1	AUGUSTUS	transcript	2651	3100	1	+	.	g4.t1
unnamed-1	AUGUSTUS	start_codon	2651	2653	.	+	0	transcript_id "g4.t1"; gene_id "g4";
unnamed-1	AUGUSTUS	single	2651	3100	1	+	0	transcript_id "g4.t1"; gene_id "g4";
unnamed-1	AUGUSTUS	CDS	2651	3100	1	+	0	transcript_id "g4.t1"; gene_id "g4";
unnamed-1	AUGUSTUS	stop_codon	3098	3100	.	+	0	transcript_id "g4.t1"; gene_id "g4";
# start gene g5
unnamed-1	AUGUSTUS	gene	3066	3941	0.68	-	.	g5
unnamed-1	AUGUSTUS	transcript	3066	3941	0.68	-	.	g5.t1
unnamed-1	AUGUSTUS	stop_codon	3066	3068	.	-	0	transcript_id "g5.t1"; gene_id "g5";
unnamed-1	AUGUSTUS	single	3066	3941	0.68	-	0	transcript_id "g5.t1"; gene_id "g5";
unnamed-1	AUGUSTUS	CDS	3066	3941	0.68	-	0	transcript_id "g5.t1"; gene_id "g5";
unnamed-1	AUGUSTUS	start_codon	3939	3941	.	-	0	transcript_id "g5.t1"; gene_id "g5";
```

==========GENEMARK==========

Model information: Staphylococcus_aureus_COL

FASTA definition line: empty-fasta-def-line
Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	267	368	102	2
2	+	559	654	96	1
3	+	1798	2313	516	1
4	+	2651	3100	450	1
5	-	3066	3941	876	1

==========ANALYSIS==========

First experiment: As a eukaryotic organism, Drosophila has one gene with multiple exons. Both AUGUSTUS and GeneMark derive almost the same result. There are three exons in gene 1: 89 to 332, 647 to 2394, 2967 to 3218 (AUGUSTUS marks the first exon from 56 to 332, with 89 as the location of start codon).

Second experiment: As a prokaryotic organism, E.coli has multiple genes. In the analyses of both AUGUSTUS and GeneMark, this whole genome shotgun sequence of E.coli has 5 genes, respectively on strand +, -, -, +, +. In AUGUSTUS, the gene locations are: 96 to 599, 707 to 2065, 2065 to 2847, 2869 to 3282, 3391 to 3660. In GeneMark, the gene locations are: 72 to 599, 707 to 2065, 2065 to 2784, 2869 to 3282, 3391 to 3660. AUGUSTUS and GeneMark do not differ much about the locations. AUGUSTUS has 3 locations in common with GeneMark, except gene 1 shorter and gene3 longer.

Third experiment: Again as a prokaryotic organism, the sequence of Staphylococcus has the same number of genes as E. coli (maybe because both sequences have similar length?). Results from AUGUSTUS and GeneMark differ especially in gene1, then getting more similar and later reaching the same in the following four genes. In AUGUSTUS, the gene locations are: 462 to 572, 574 to 654, 1798 to 2313, 2651 to 3100, 3066 to 3941. In GeneMark, the gene locations are: 267 to 368, 559 to 654, 1798 to 2313, 2651 to 3100, 3066 to 3941.

This similarity between the two methods may be caused by the fact that both methods use hidden markov models with Viterbi's algorithm. Other shared characteristics also include that they use triplets as a state in their markov chain, that they import interpolation (e.g. choosing prior

positions by Chi-sqare test, variance order, and delete interpolation that combines higher and lower models) in their markov chain instead of fixed order (FO), that they implement forward-backward algorithm to consider both states ahead of the current state (left side of the current position in sequence) and states behind the current state (right side of the current position).

But there are differences as well between AUGUSTUS and GeneMark, which do not affect much given the input. First, AUGUSTUS can take external hints and constraints from EST, mRNA, or protein databases, retrieve them by BLAST, integrate them with intrinsic information, then combine them to the model; whereas GeneMark is merely based on the sequence. However, since our input do not give the external hints, this difference does not make a change. Furthermore, GeneMark implements Bayesian formalism, which functions similarly to forward-backward algorithm yet baysian probability is more generalized. GeneMark makes use of sliding window to handle sparse distribution of genes/exons, whose purpose is to reduce the complexity instead of making a significant difference in result. GeneMark also embeds maximum likelyhood in parsing DNA, which again simplifies the calcuation without affecting the result. For AUGUSTUS it implements windowed weight array model (WWAM) with similarity-based weighting of sequence patterns. This feature is useful for AUGUSTUS to analyse the distribution of emission, pattern length, and sequence. Yet, because ab initio approach is statistically similarity based, this feature does not make a great difference.