Assignment 4 Extra credit

Krishan Subudhi (ksubudhi@uw.edu)

Student number: 2040900

7. predicted RNAs that are not annotated in the GFF file

Another annotation source:

Along with NCBI I found another source for the genome which was updated recently.

ftp://ftp.ensemblgenomes.org/pub/bacteria/release-48/gff3/bacteria_0_collection/methanocaldococcus_jannaschii_dsm_2661

Methanocaldococcus_jannaschii_dsm_2661.ASM9166v1.46.gff3.gz

Source : http://bacteria.ensembl.org/Methanocaldococcus_jannaschii_dsm_2661/Info/Index

	state	start	end	length	percentage_overlap	percentage_overlap2
0	State 2	97326	97541	216	51.851852	51.851852
1	State 2	97627	97823	197	44.670051	44.670051
2	State 2	111764	111856	93	95.698925	95.698925
3	State 2	118079	118179	101	0.000000	0.000000
4	State 2	138345	138419	75	100.000000	100.000000
5	State 2	154610	157697	3088	96.437824	96.437824
6	State 2	157782	159591	1810	85.801105	85.801105
7	State 2	186974	187067	94	94.680851	94.680851
8	State 2	190831	190907	77	98.701299	98.701299
9	State 2	215200	215296	97	89.690722	89.690722
10	State 2	227705	227782	78	97.435897	97.435897
11	State 2	291972	291997	26	0.000000	0.000000
12	State 2	303990	304080	91	97.802198	97.802198
13	State 2	358766	358942	177	85.875706	85.875706
14	State 2	359974	360046	73	100.000000	100.000000
15	State 2	402969	403057	89	85.393258	85.393258
16	State 2	412582	412635	54	0.000000	0.000000
17	State 2	552537	552862	326	96.932515	96.932515

	state	start	end	length	percentage_overlap	percentage_overlap2
18	State 2	619161	619236	76	97.368421	97.368421
19	State 2	637579	638153	575	81.043478	81.043478
20	State 2	638334	640132	1799	86.492496	86.492496
21	State 2	640217	643449	3233	95.731519	95.731519
22	State 2	643500	643767	268	96.268657	96.268657
23	State 2	763767	763845	79	96.202532	96.202532
24	State 2	764022	764095	74	100.000000	100.000000
25	State 2	774708	774788	81	0.000000	0.000000
26	State 2	863476	864151	676	73.816568	73.816568
27	State 2	873579	873778	200	39.000000	39.000000
28	State 2	883675	883755	81	91.358025	91.358025
29	State 2	951852	951968	117	0.000000	0.000000
30	State 2	1038544	1038622	79	97.468354	97.468354
31	State 2	1129124	1129194	71	0.000000	0.000000
32	State 2	1150142	1150402	261	58.237548	58.237548
33	State 2	1189943	1190054	112	98.214286	98.214286
34	State 2	1313165	1313251	87	97.701149	97.701149
35	State 2	1659451	1659520	70	0.000000	0.000000

It looks like both the sources (NCBI, bacteria.ensemble.org) point to the same annotation

1. RFAM showed the tRNAs (generally very small sequences) but there was no way to download them. From a brief overview it looks like, the website does not have much information.
2. RNACentral lets you download all the searches but the downloaded file does not have position information in it.

```
df[df.percentage_overlap <50]
```

	state	start	end	length	percentage_overlap	percentage_overlap2
1	State 2	97627	97823	197	44.670051	44.670051
3	State 2	118079	118179	101	0.000000	0.000000
11	State 2	291972	291997	26	0.000000	0.000000
16	State 2	412582	412635	54	0.000000	0.000000

	state	start	end	length	percentage_overlap	percentage_overlap2
25	State 2	774708	774788	81	0.000000	0.000000
27	State 2	873579	873778	200	39.000000	39.000000
29	State 2	951852	951968	117	0.000000	0.000000
31	State 2	1129124	1129194	71	0.000000	0.000000
35	State 2	1659451	1659520	70	0.000000	0.000000

Except for 27, none of the predictions are very long. So the model is not missing major genes.

8. Evaluation Metric

Q. Do a more formal analysis of the "accuracy" of this method for its designated purpose of discovering RNA genes. I.e., count True Positives, False Negatives, etc.

Based on the paper, *Assessing computational tools for the discovery of transcription factor binding sites* I have calculate **Specificity and PPV metrics** at neucleotite levels.

- nTP is the number of nucleotide positions in both known sites and predicted sites,
- nFN is the number of nucleotide positions in known sites but not in predicted sites,
- nFP is the number of nucleotide positions not in known sites but in predicted sites, and
- nTN is the number of nucleotide positions in neither known sites nor predicted sites.

```
df['overlap_length'] = overlap_lengths
df.head()
```

	state	start	end	length	percentage_overlap	percentage_overlap2	overlap_length
0	State 2	97326	97541	216	51.851852	51.851852	112
1	State 2	97627	97823	197	44.670051	44.670051	88
2	State 2	111764	111856	93	95.698925	95.698925	89
3	State 2	118079	118179	101	0.000000	0.000000	0
4	State 2	138345	138419	75	100.000000	100.000000	75

```
genepos2.head()
```

	start	end	length
0	97426	97537	111
1	97629	97716	87
2	111766	111854	88
3	138344	138419	75
4	154662	157639	2977

```
nTP = df.overlap_length.sum()
nFN = genepos2.length.sum() - nTP
nFP = df.length.sum() - nTP
nTN = 1664970 - (nTP+nFP+nFN)

nTP, nFN, nFP, nTN
```

```
(12603, 42, 2098, 1650227)
```

Sensitivity

Sensitivity gives fraction of known site neucleotites that are predicted

Sensitivity: nSn = nTP/(nTP + nFN), and

```
nSn = nTP/(nTP + nFN)
print('Sensitivity  = ', nSn)
```

```
Sensitivity  =  0.9966785290628707
```

PPV

PPV gives fraction of predicted site neucleotites that are known

Positive Predictive Value: nPPV = nTP/(nTP + nFP)

```
nPPV =  nTP/(nTP + nFP)
print('PPV  = ', nPPV)
```

```
PPV  =  0.8572886198217808
```

Many predictions are either starting before the gene starts or still continue after the gene ends. Because of to many false positives, the PPV value is low for this model.

Note on PPV and Sensitivity:

For datasets with no binding sites (TP + FN = 0) , sensitivity is not defined while PPV is uninformative.

If the model predicts no hit state (TP + FP = 0) , PPV is undefined and sensitivity is undefined.