Sagex Error Analysis

W. Evan Durno

November 26, 2014

General Design

To analyze Sagex for classification errors, we take three SAGs and download their respective genomes from NCBI. The SAGs are from samples which also have metagenomes From the genome, we subset with replacement 2000 reads of 3000 bp length and append the subset and the three metagenomes together to form a labelled metagenome. Sequences from the genome are obviously labelled as correct hits in extrapolation. Sequences from the metagenome are naively assumed to be false positives, though this is obviously not always true. In all, Sagex is shown to be a powerful classifier.

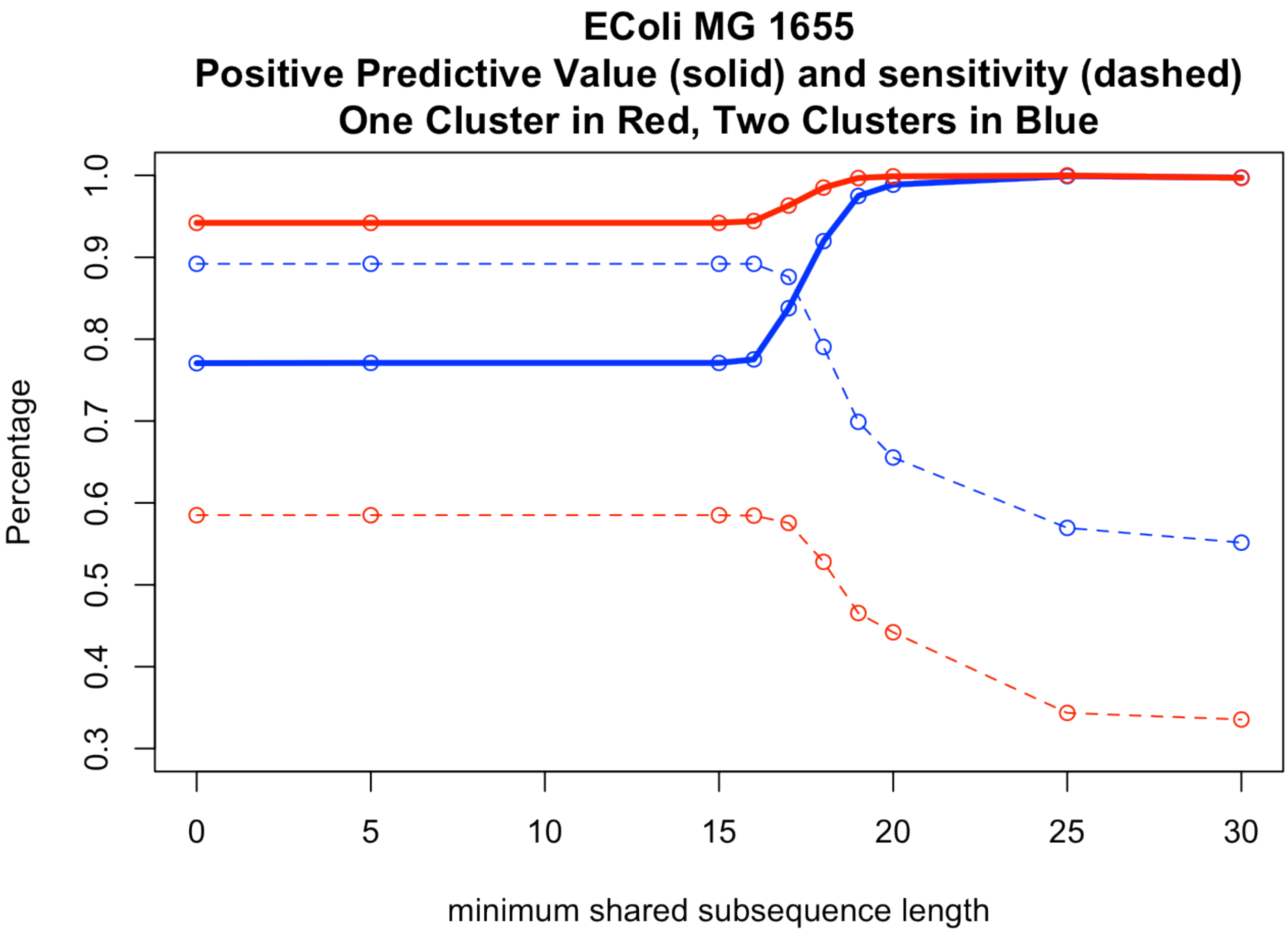
```
# Example usage:
./sagex -i sag.fasta -G metaGenomeAndGenomeSubset.fasta -v -P -k 2 -C 20 > output
```

EColi MG1655

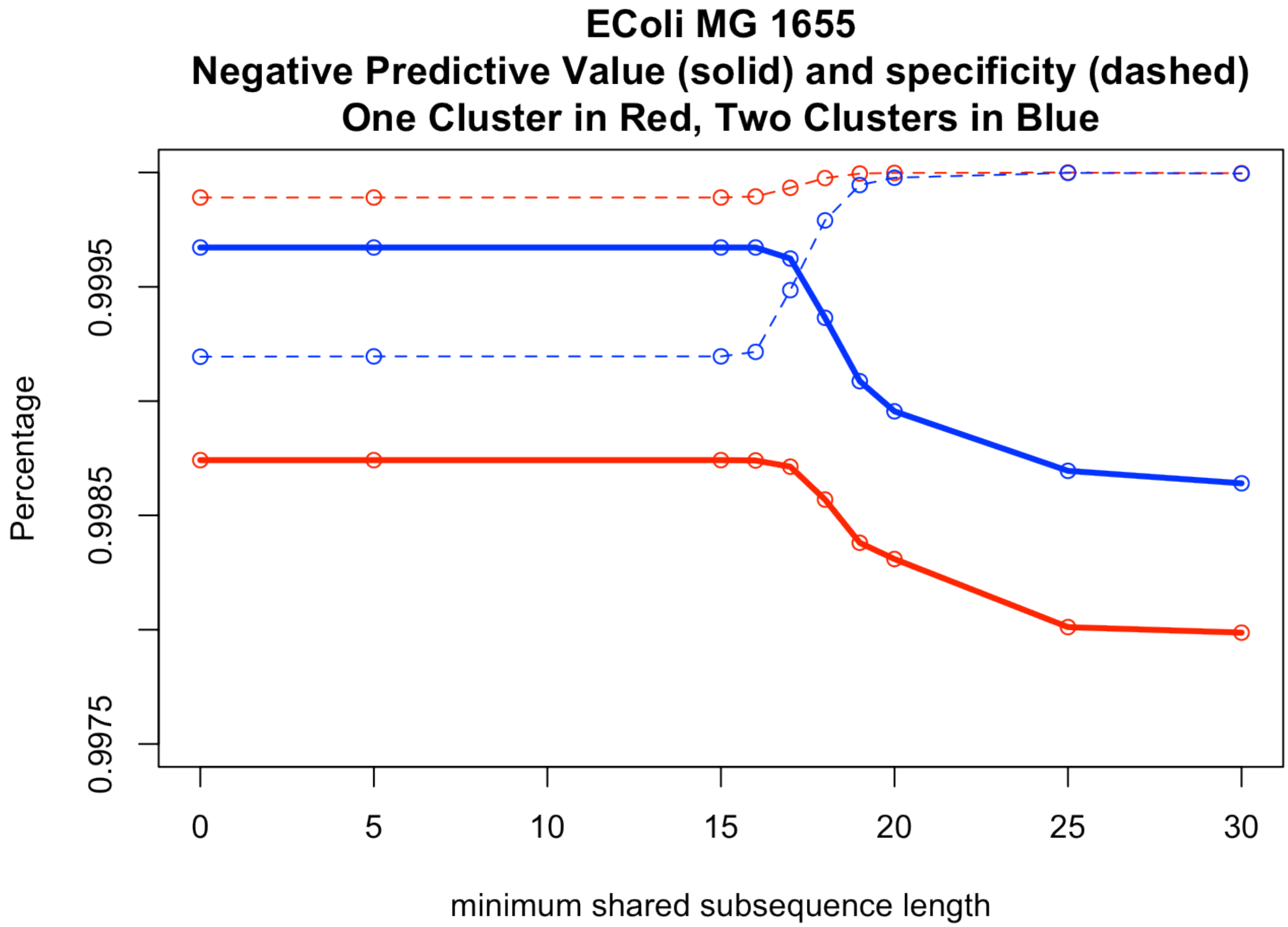
This SAG analyzes function under perfect conditions. This SAG was generated to validate a microfluidics tool. There is minimal opportunity for genetic drift between SAG and genome.

```
kVec = c( 0 , 5 , 15 , 16 , 17 , 18 , 19 , 20 , 25 , 30 )
correctHits1 = c( 1170 , 1170 , 1170 , 1169 , 1151 , 1056 , 931 , 884 , 687 , 671 )
hits1 = c( 1242 , 1242 , 1242 , 1238 , 1195 , 1072 , 934 , 885 , 687 , 673 )
correctHits2 = c( 1784 , 1784 , 1784 , 1784 , 1752 , 1581 , 1398 , 1311 , 1139 , 1103 )
hits2 = c( 2315 , 2314 , 2314 , 2301 , 2091 , 1719 , 1434 , 1326 , 1140 , 1106 )
falsePositives1 = hits1 - correctHits1
falsePositives2 = hits2 - correctHits2
trueNegatives1 = 658928 - falsePositives1
trueNegatives2 = 658928 - falsePositives2
falseNegatives1 = 2000 - correctHits1
falseNegatives2 = 2000 - correctHits2

plot( kVec , correctHits2 / hits2 , ylab="Percentage" , xlab="minimum shared subsequence length" , main="EColi MG 1655\nPositive Predictive Value (solid) and sensitivity (dashed)\nOne Cluster in Red, Two Clusters in Blue" , col="blue" , ylim = c(0.3,1.0) )
lines( kVec , correctHits2 / hits2 , type ='l' , lwd = 3 , col="blue" )
points( kVec , correctHits1 / hits1 , col="red" )
lines( kVec , correctHits1 / hits1 , type ='l' , lwd = 3 , col="red" )
lines( kVec , correctHits2 / 2000 , lty=2 , type="o" , col = "blue" )
lines( kVec , correctHits1 / 2000 , lty=2 , type="o" , col = "red" )
```



```
plot( kVec , trueNegatives1 / (trueNegatives1 + falseNegatives1) , ylab="Percentage" , xlab="minimum shared subsequence length" , main="EColi MG 1655\nNegative Predictive Value (solid) and specificity (dashed)\nOne Cluster in Red, Two Clusters in Blue" , col="red" , ylim=c( 0.9975,1 ) )
lines( kVec , trueNegatives1 / (trueNegatives1 + falseNegatives1) , lwd = 3 , col="red" )
points( kVec , trueNegatives2 / (trueNegatives2 + falseNegatives2) , col="blue" )
lines( kVec , trueNegatives2 / (trueNegatives2 + falseNegatives2) , lwd = 3 , col="blue" )
lines( kVec , trueNegatives1 / 658928 , lty=2 , type="o" , col="red" )
lines( kVec , trueNegatives2 / 658928 , lty=2 , type="o" , col="blue" )
```



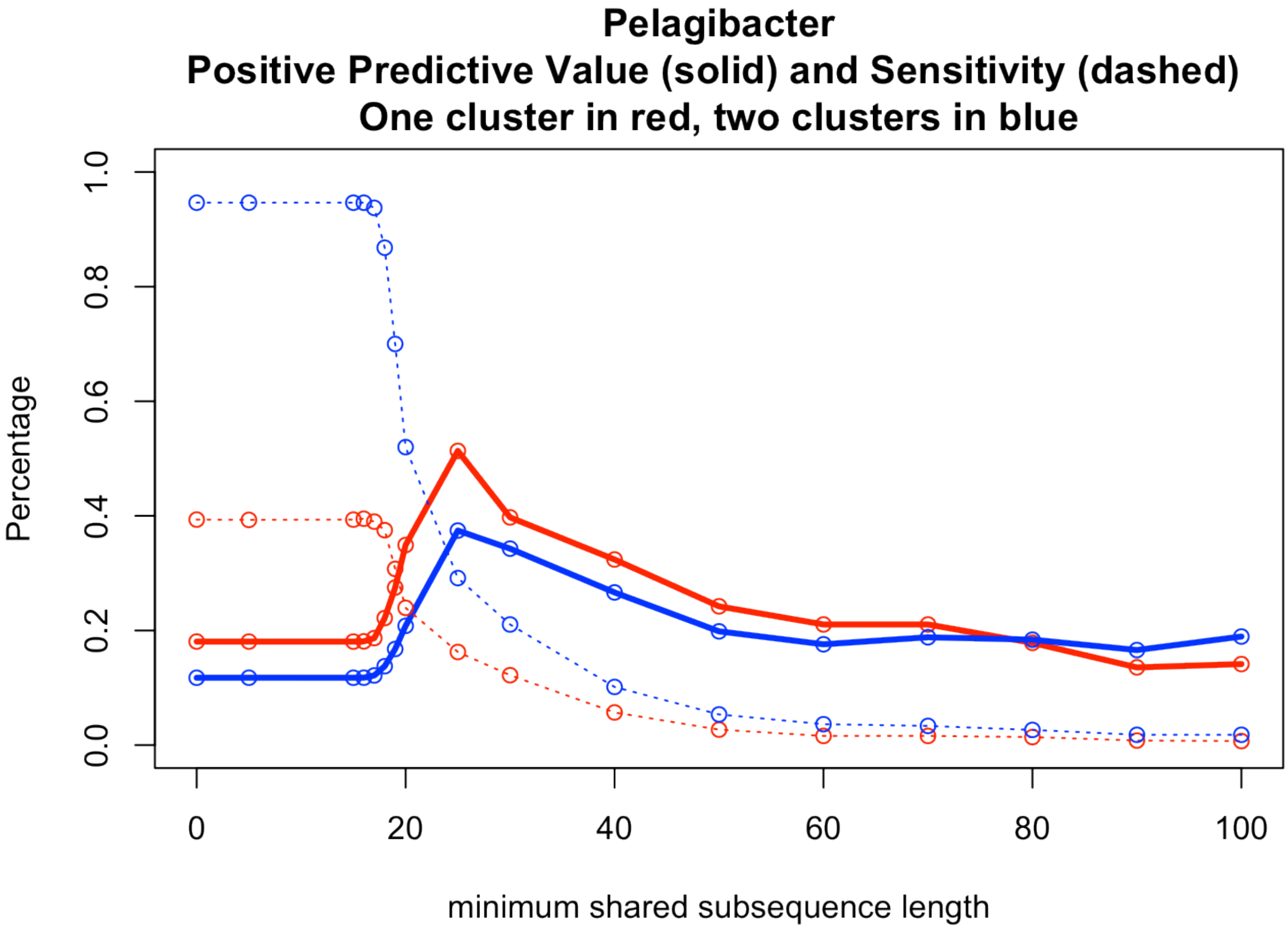
SAR11

This SAG is genetically distant from its genome and thus, as classifier stringency is increased, Sagex begins siding with the the metagenome over the SAG. Superficially, this makes the errors appear to worsen, but it merely demonstrates how different the genome is from the SAG at the genetic level.

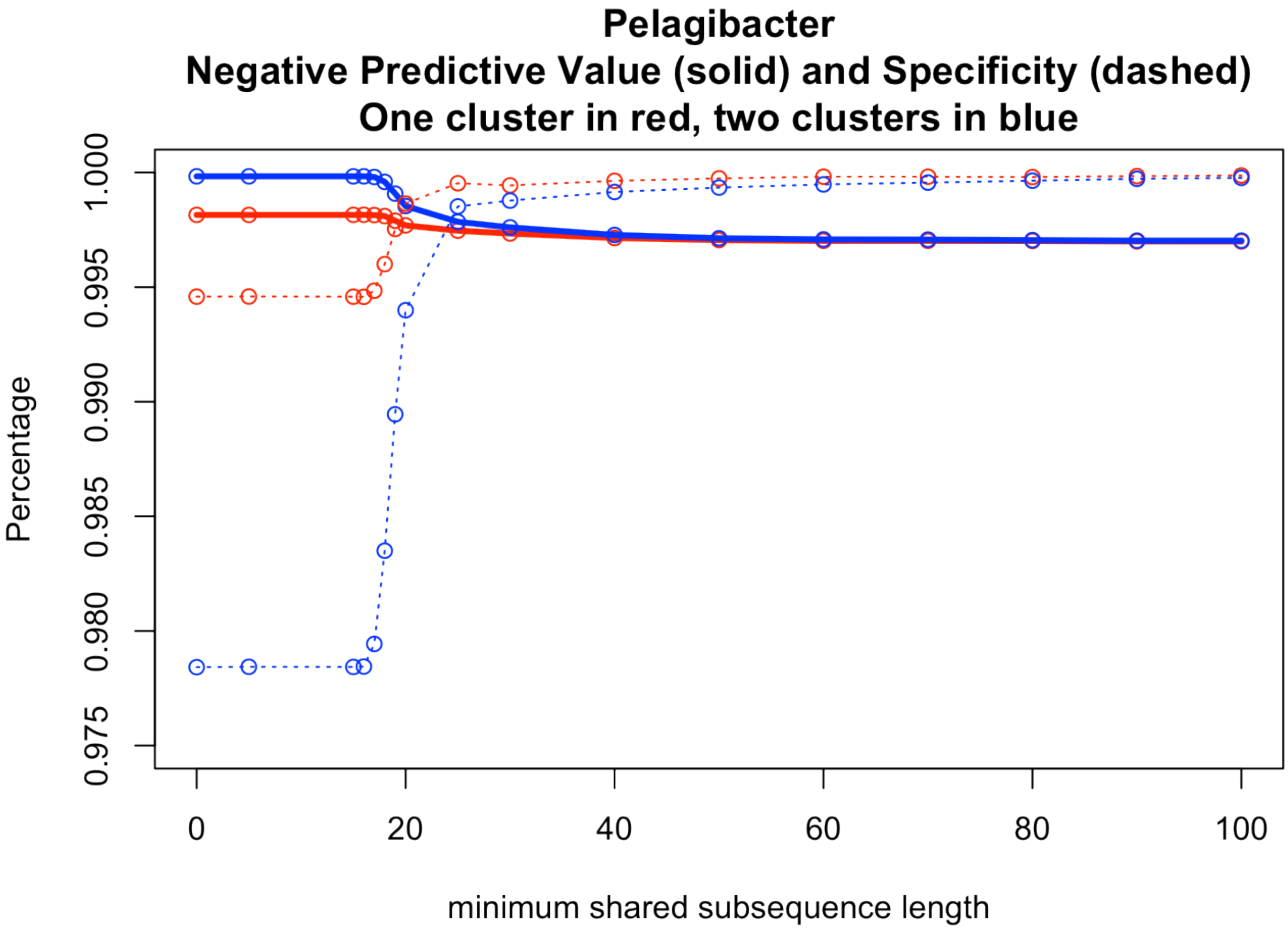
```
kVec = c( 0 , 5 , 15 , 16 , 17 , 18 , 19 , 20 ,25 , 30 , 40 , 50 , 60 , 70 , 80 , 90 , 100 )
trueHits1 = c( 787 , 786 , 787 , 790 , 780 , 750 , 615 , 479 , 325 , 244 , 114 , 54 , 32 ,32 , 28 , 16 , 14 )
trueHits2 = c( 1893 , 1893 , 1893 , 1893 , 1875 , 1736 , 1400 , 1040 , 583 , 421 , 203 , 107 , 73 , 67 , 53 , 36 , 36 )
hits1 = c( 4355 , 4350 , 4356 , 4364 , 4178 , 3383 , 2237 , 1372 , 633 , 614 , 352 , 223 , 152 , 152 , 157 , 118 , 99 )
hits2 = c( 16113 , 16102 , 16105 , 16094 , 15426 , 12608 , 8350 , 4999 , 1557 , 1228 , 762 , 539 , 415 , 356 , 288 , 217 , 190 )
falsePositives1 = hits1 - trueHits1
falsePositives2 = hits2 - trueHits2
```

```
trueNegatives1 = 658928 - falsePositives1
trueNegatives2 = 658928 - falsePositives2
falseNegatives1 = 2000 - trueHits1
falseNegatives2 = 2000 - trueHits2

plot( kVec , trueHits1 / hits1 , ylab="Percentage" , xlab="minimum shared subsequence length" , main="Pelagibacter\nPositive Predictive Value (solid) and Sensitivity (dashed) \nOne cluster in red, two clusters in blue" , col="red" , ylim=c(0,1) )
lines( kVec , trueHits1 / hits1 , lwd=3 , col="red" )
points( kVec , trueHits2 / hits2 , col="blue" )
lines( kVec , trueHits2 / hits2 , lwd=3 , col="blue" )
lines( kVec , trueHits1 / 2000 , lty=3 , type="o" , col="red" )
lines( kVec , trueHits2 / 2000  , lty=3 , type="o" , col="blue" )
```



```
plot( kVec , trueNegatives1 / (trueNegatives1 + falseNegatives1) , ylab="Percentage" , xlab="minimum shared subsequence length" , main="Pelagibacter\nNegative Predictive Value (solid) and Specificity (dashed)\nOne cluster in red, two clusters in blue" , col="red" , ylim=c(
0.975,1) )
lines( kVec , trueNegatives1 / (trueNegatives1 + falseNegatives1) , lwd=3 , col="red" )
points( kVec , trueNegatives2 / (trueNegatives2 + falseNegatives2) , col="blue" )
lines( kVec , trueNegatives2 / (trueNegatives2 + falseNegatives2) , lwd=3 , col="blue" )
lines( kVec , trueNegatives1 / 658928 , lty=3 , type="o" , col="red" )
lines( kVec , trueNegatives2 / 658928 , lty=3 , type="o" , col="blue" )
```

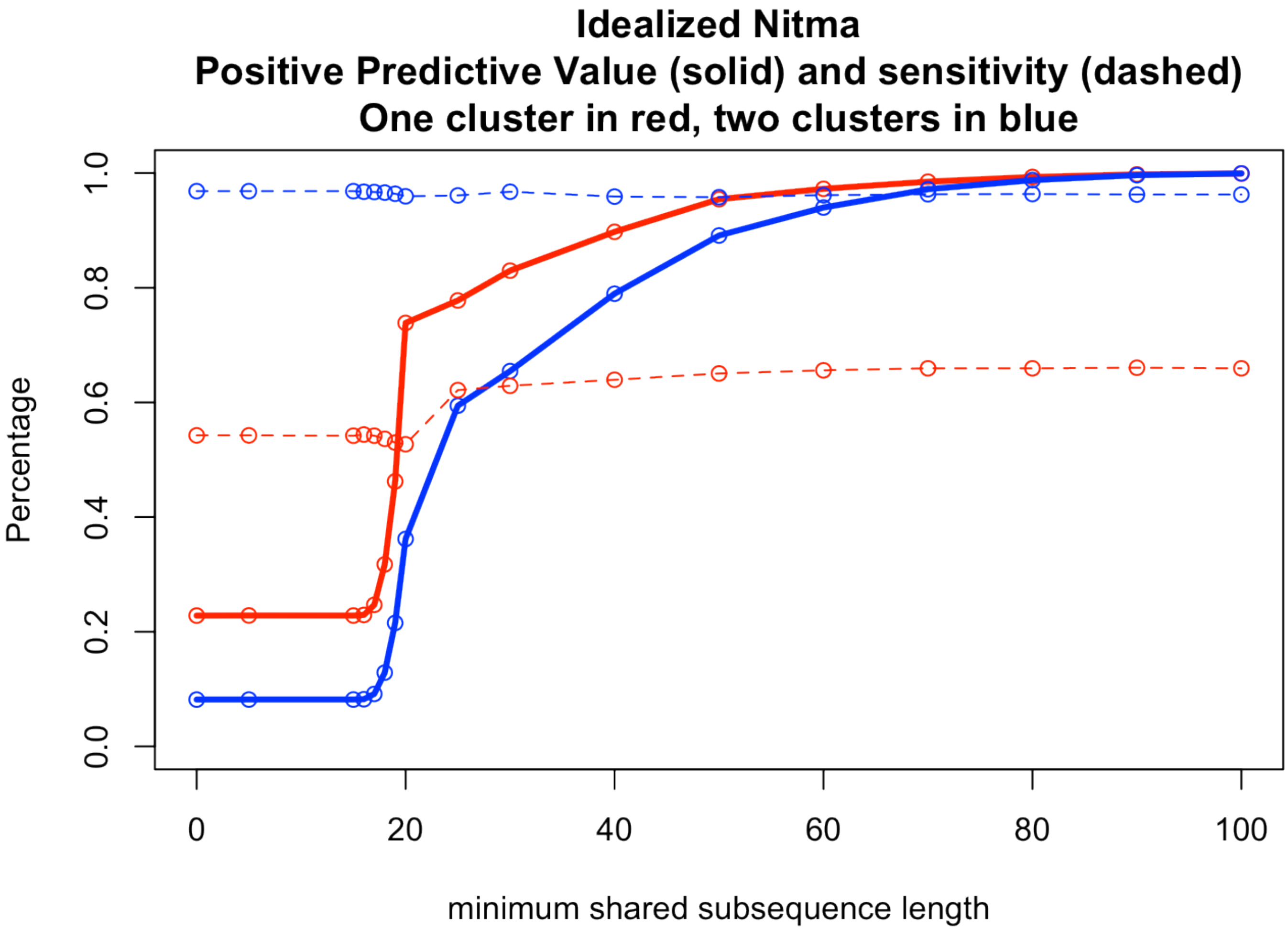


Idealized Nitma SAG

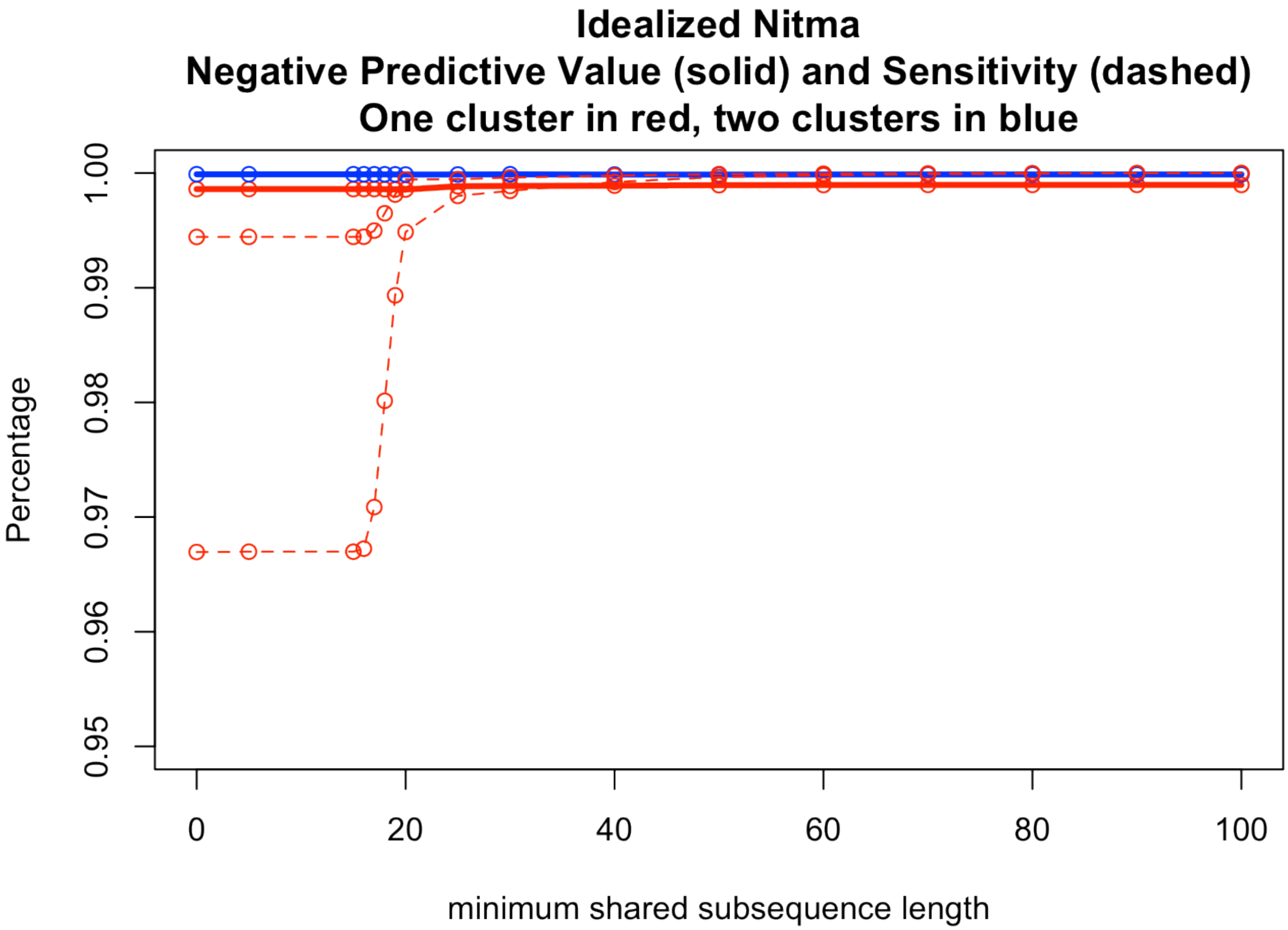
No Nitma genome exists in our set of SAGs, but it does exist abundantly within the metagenome. For a SAG, we randomly simulate 2000 reads of length 3000 bp from a known Nitma genome. This analysis demonstrates how classification errors appear extremely favourable with a sufficiently stringent cutoff.

```
kVec = c( 0 , 5 , 15 , 16 , 17 , 18 , 19 , 20 , 25 , 30 , 40 , 50 , 60 , 70 , 80 ,90 , 100
)
trueHits1 = c( 1085 , 1085 , 1084 , 1088 , 1084 , 1073 , 1060 , 1054 , 1243 , 1258 , 1279 ,
1301 , 1312 , 1319 , 1319 , 1321 , 1319 )
trueHits2 = c( 1937 , 1937 , 1937 , 1935 , 1934 , 1932 , 1928 , 1919 , 1922 , 1935 , 1918 ,
1916 , 1923 , 1926 , 1927 , 1925 , 1925 )
hits1 = c( 4755 , 4750 , 4749 , 4745 , 4389 , 3380 , 2293 , 1427 , 1598 , 1516 , 1425 , 136
3 , 1349 , 1339 , 1328 , 1324 , 1320 )
hits2 = c( 23711 , 23695 , 23694 , 23518 , 21127 , 15017 , 8953 , 5303 , 3234 , 2956 , 2429
, 2150 , 2046 , 1982 , 1951 , 1932 , 1926 )
falsePositives1 = hits1 - trueHits1
falsePositives2 = hits2 - trueHits2
trueNegatives1 = 658928 - falsePositives1
trueNegatives2 = 658928 - falsePositives2
falseNegatives1 = 2000 - trueHits1
falseNegatives2 = 2000 - trueHits2


plot( kVec , trueHits1 / hits1 , ylab="Percentage" , xlab="minimum shared subsequence lengt
h" , main="Idealized Nitma\nPositive Predictive Value (solid) and sensitivity (dashed)\nOne
cluster in red, two clusters in blue" , col="red" , ylim=c(0,1) )
lines( kVec , trueHits1 / hits1 , lwd=3 , col="red" )
points( kVec , trueHits2 / hits2 , col="blue" )
lines( kVec , trueHits2 / hits2 , lwd=3 , col="blue" )
lines( kVec , trueHits1 / 2000 , lty=2 , type="o" , col="red" )
lines( kVec , trueHits2 / 2000 , lty=2 , type="o" , col="blue" )
```



```
plot( kVec , trueNegatives1/(trueNegatives1 + falseNegatives1) , ylab="Percentage" , xlab="minimum shared subsequence length" , main="Idealized Nitma\nNegative Predictive Value (solid) and Sensitivity (dashed)\nOne cluster in red, two clusters in blue" , col="red" , ylim=c(0.95,1) )
lines( kVec , trueNegatives1/(trueNegatives1 + falseNegatives1) , lwd=3 , col="red" )
points( kVec , trueNegatives2/(trueNegatives2 + falseNegatives2) , col="blue" )
lines( kVec ,trueNegatives2/(trueNegatives2 + falseNegatives2) , lwd=3 , col="blue" )
lines( kVec , trueNegatives1/658928 , lty=2 , type="o" , col="red" )
lines( kVec , trueNegatives2/658928 , lty=2 , type="o" , col="red" )
```


Idealized Nitma
Negative Predictive Value (solid) and Sensitivity (dashed)
One cluster in red, two clusters in blue
Percentage
0.95
0.96
0.97
0.98
0.99
1.00
0
20
40
60
80
100
minimum shared subsequence length