Random Forest Classifier Based Approach for Next Generation Sequencing Gene Data Classification

A. Q. M. Sala Uddin Pathan, Md Hasnat Riaz, Mohammad Humayun Kabir, Md Javed Hossain

Abstract— Next Generation Sequencing (NGS) opens a new door to the researcher by permits high throughput of gene data sets at low cost with high accuracy. This analysis produces a massive amount of biosequence which required enormous storage capacity. Several methods have been proposed in terms of assembly and alignment to limit the storage problem and classify the pathogen. In this paper, for classifying specific pathogen, we described a Random Forest based alignment-free approach that converts the base sequence into different lengths k-mers. This method can be used for classification of an organism. We have shown the usefulness of this approach for the analysis of Staphylococcus Aureus. We were able to identify unique features that were used for the classification.

Index Terms— Next Generation Sequencing, Alignment-free approach, Gene data Classification, Random Forest, k-mer, low cost, high accuracy

1 Introduction

Next Generation Sequencing (NGS) [1], build on a traditional shotgun sequencing, is a combination of various DNA sequencing technologies. In laboratory sequencing machines arbitrarily break DNA up into many short known as reads. The resulting data are stored in the manufacturing standard FASTQ structure. Because of computationally rigorous alignment steps, the amount of data created is quickly outperforming analytics abilities. For quick diagnosis of the bacterial infection, the need for identifying the species is very important. Based on this alignment-free approach is a good solution to this.

Alignment-free approach haste up the reading as well as delivers a way to gain useful gene data from raw shorts. K-mers are built from the shorts and used to calculate the reads by matching them to each sequencing read directly. Based on these, RF (Random Forest) used to classify whether the entire sequencing project belongs to a particular organism. Earlier many classifier algorithms have been used for this purpose.

Bhuvaneswari et al. [3] projected an outline to expose informative gene sequences and to classify gene patterns fitting to its related subtype by using fuzzy logic. Fuzzy systems adjust numerical data into human phonological terms, which propose very decent capabilities to deal with noisy and missing data. Yet, defining the rules and membership functions needs a lot of prior information from human professional.

Due to its capability to plot the input-output data, Artificial Neural Networks (ANN) has been proposed for gene sequence data classification. Bevilacqua et al. [4] used Feed-Forward NN to develop a perfect classifier. Khan et al. [5] used neural networks to examine microarray data from patients with slight round blue-cell tumors. In Chen et al. [6] presented categorizing gene sequence data using artificial neural network ensembles established on samples filtering.

Neural networks can plot the input data into various classes directly with a single network. Besides, the neural network systems can easily adapt nonlinear features of the gene sequence data [5]. Neural networks can also be easily adjusted to yield continuous variables instead of discrete class markers. This will be useful for situations where we want to predict the level of the medical pointer rather than classify the samples into binary categories [7]. But the Neural networks generally accept gradient-based learning approaches, which are vulnerable to local minima and require a long time for training [8].

Recently, Support Vector Machines have been suggested for gene data classification in the gene classification approach presented by Furey et al. [9]. SVM classification is typically parallelizable, but efforts to parallelize the training process have recognized far less successful.

The comparative complexity and consecutive structure of SVM training, we pursue to use methods that may be more efficiently used in systems which partition the training and classification problems to seize advantage of the accessible computational resources.

Random Forests are ensemble approaches which use a number of independent decision trees for classification. Each tree uses a random selection of features to craft a classification decision, and the Random Forest categorizes by selecting the mode of all the decision tree outputs. By using a large number of decision trees, a Random Forest is able to overcome the difficulties associated with using a random selection of features to make a decision. Decision Trees are well-established classification methods that recursively fragmented a data set into smaller sets based on the effect of a test described at each branch in the tree [2]. Starting with the root node, decisions are made till a leaf node is reached, at which fact the suitable label can be applied to the data point. The challenge in the training process is to decide which feature to use at each decision to split the data points. A Random Forest beats this semi-random feature selection by forming a large number of decision trees, in the hundreds or even thousands. Each tree can be built independently, which creates Random Forest structure MapReduceable, with the drop of step incorporating an overall decision on the classification of the example data vector. Our main objective of this work is to classify whether the entire sequencing project belongs to a particular organism.

Figure 1.1 Scanning electron micrograph of S. aureus.

For our work, we chose Staphylococcus Aureus as a classification organism. S. Aureus is a familiar fellow of the microbiota of the body, commonly originate in the upper respiratory area and on the skin. It causes food poisoning, respiratory infections, and skin infections containing swellings. Almost 20% to 30% of the humen are a carrier of this bacteria. [13] [14] In spite of widely research, no vaccine has been approved for it.

2 METHODOLOGY

Analysis of next-generation sequencing (NGS) records set a huge challenge. It needed a methodical and smart approach to process the NGS data expertly. The first thing to do with sequencing details is to compute the value of sequencing data. For example, we will acquire a general view on the number and length of reads, if there are any polluting sequences in the sample or low-quality sequences. After examining the quality of the data and if necessary, pre-processed it, the next phase is mapping, also called aligning. It allows controlling the nucleotide sequence of records being reviewed with no necessity of de novo assembly as acquired reads are compared with a reference previously been in a database. Once plotting the reads, it is a good idea to test the mapping excellence, as some of the biases in the record only highlight after the mapping phase. Once the sequence is aligned to a reference genome, we can predict from the difference whether the entire sequencing belongs to a particular organism.

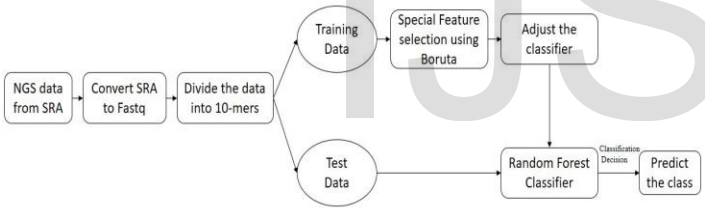

Figure 2.1 Schematic diagram of the classification approach

In this work, we chose data set over a blend of data: real (pulled from the Sequence Read Archive (SRA) [10]. Then we converted SRA data of each project to industry standard FASTQ format by using SRA toolkit [11].

A sliding window of length k was used to count the occurrences of each k -mer across each read. These counts were then accumulated over all the reads in the project and then normalized, resulting in a feature vector of substantial length for each project. For k=10, these feature vectors were reduced using the Boruta [12] feature selection algorithm, as not all features were significant in classification. From each project, we took 15000 features to evaluate the data. Boruta, a wrapper based algorithm, tries to use a subset of features and train a model using them. Based on the inferences one can decide to add or remove features from the subset.

Two classes were used for classification 1 and 0. Here 1 represents S. aureus and 0 represents not S. aureus. The classifier was able to distinguish Staphylococcus Aureus successfully. As noted above, concerns over the scalability of this approach led us to consider a collection of simpler or at least more simply trainable classifier elements to allow parallelization.

In the trial studies using the R caret package, we found that RF with regularization outperformed all other classifiers. We split the projects into train and test sets consisting of 70% and 30% projects respectively. 10-fold cross-validation was used on the training set to calculate mean accuracy and standard deviation, while the holdout set was used to calculate holdout accuracy, precision, and recall. Trials using k=10 yielded an optimal selection of features.

```
# K-mer the sequence for certain length
find_kmers <- function(string, k){
  n <- nchar(string) - k + 1
  kmers <- substring(string, 1:n, 1:n + k - 1)
  return(kmers)
}
kseq <- find_kmers(sread(fastq),10)
kseq
```

Figure 2.2 R query to 10-mer the sequence

```
# Data Partition
set.seed(222)
ind <- sample(2, nrow(h), replace = T, prob = c(0.7, 0.3))
train <- Alldata[ind==1,]
test <- Alldata[ind==2,]
```

Figure 2.3 R query to split the project into train and test sets

```
# Feature Selection
set.seed(111)
boruta <- Boruta(Class ~ ., data = Alldata, doTrace = 2, maxRuns = 1500)
print(boruta)
plot(boruta, las = 2, cex.axis = 0.1)
plotImpHistory(boruta)
```

Figure 2.4 R query to select special features

```
set.seed(2348)
cv.10.folds <- createMultiFolds(rf.label, k=10, times = 10)
#set up caret's trainControl object per above
ctrl.1 <-  trainControl(method = "repeatedcv", number = 10, repeats = 10,
                        index = cv.10.folds)
cl <- makeCluster(6, type = "SOCK")
registerDoSNOW(cl)

#Set seed for reproducibility and train
set.seed(34324)
rf.1.cv.1 <- train(x= rf.train.1, y = rf.label, method= "rf", tuneLength = 3,
                   ntree = 1000, trControl = ctrl.1)
#Shutdown cluster
stopCluster(cl)
```

Figure 2.5 R query to 10-fold cross-validation

```
#Make predictions
rf.1.preds <- predict(rf.1, test.df)
table(rf.1.preds)
```

Figure 2.6 R query to predict the sample

It is important to note that past a certain point increasing the read coverage does not improve the classification results. From experimentation, it was found that coverage of 10 was sufficient. As the k-mer counts are normalized, only the relative frequencies of each k-mer are considered when performing classification. Once the coverage is high enough for sequencing error to be normalized out, further increases only increase the time taken to prepare the data set; they do not improve the classifier performance.

3 Result and Discussion

We demonstrate efficient use of next-generation sequencing (NGS) data to predict an organism. Our study displayed that the class of an organism can be precisely predicted using the alignment-free approach with dissimilarity and that the system proposed a simple and unified approach for organism classification that can be used to any biological sample.

With the Random Forest able to effectively categorize 1 and 0 classes, it is of interest to decide whether the k-mer representation and Random Forest are capable of categorising other organisms. As above, k =10 was used for k-mer tallying, and Boruta was used on 35 projects to select the top features. 10-fold cross-validation was used on the training/test set to decide the mean accuracy and standard deviation. Top features were randomly chosen for consideration per node, and 8000 trees were structured.

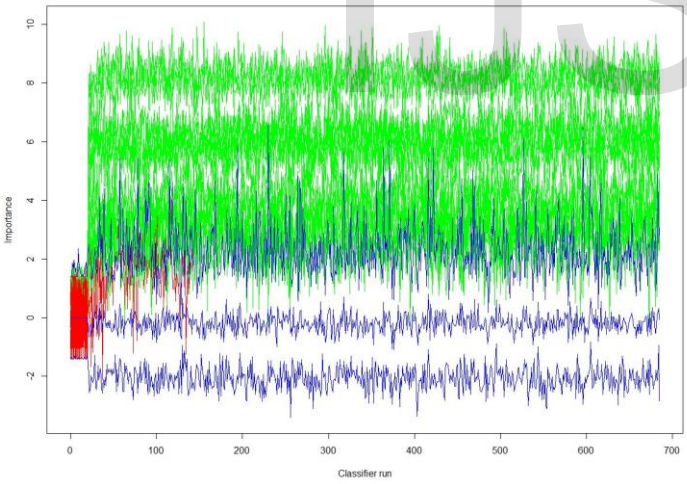

Figure 3.1 Importance of features selected by boruta

The data preparation procedure explained above was presented to create data sets of various sizes, primarily to test the computational consequences of scaling to big data sets. There was no extra substitution performed. An approximate 70/30 split was used, as this attained the best results in earlier work. 10-fold cross-validation was used on the train set to determine the mean accuracy and standard deviation of the Random Forest. The Random Forest randomly selected top features for consideration per node, and 8000 trees were constructed. Random Forest was able to perfectly classify all projects at various data set sizes, albeit trivially as the noise level was zero. The timings for data preparation are provided above, and the training and classification timings were trivial in comparison.

4 Conclusion

The analysis of enormous sequencing data from next-generation sequencer needs ample storage, main memory space, and it requires much time. In this paper, we have endeavored to overcome these problems using a new approach to transform a fastq to k-mer size and run a feature selection algorithm for finding special features. We run random forest classifier on train data to test whether the sequence belongs to a specific organism. It successfully classified the corresponding organism.

References

[1] M. L. Metzker, "Sequencing technologies - the next generation," Nature Reviews Genetics, vol. 11, pp. 31-46, 2010.

[2] M. Friedl and C. Bradley, "Decision tree classification of land cover from remotely sensed data," ” Remote sensing of environment, vol. 61, no. 3, pp. 399-409, 1997.

[3] Bhuvaneswari V, Vanitha K. "Classification of Microarray Gene Expression Data by Gene Combinations using Fuzzy Logic." International Journal of Computer Science, Engineering and Applications 2012; 2 -4, p.79 – 98.

[4] Bevilacqua V, Mastronardi G, Menolascina F, Paradiso A, Tommasi S. "Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis¬: a Distributed Approach" 2006; 14(November).

[5] Khan J, Wei,M.Ringner JS, Saal,M.Ladanyi LH, Westermann,F.Berthold F, Schwab M, Antonescu CR, Peterson C, and Meltzer S, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks." Nature Medicine 2001; 7 -6, p. 673 -679.

[6] Wutao Chen, Huijuan Lu, Mingyi Wang, and Cheng Fang. "Gene Expression Data Classification Using Artificial Neural Network Ensembles Based on Samples Filtering." International Conference on Artificial Intelligence and Computational Intelligence 2009; 1, p. 626 – 628.

[7] Ringner M, Peterson C, Khan J. "Analyzing Array Data Using Supervised Methods. "Pharmacogenomics 2002; 3- 3, p. 403 -415.

[8] Linder R, Dew D, Sudhoff H, Theegarten D, Remberger K, Poppl SJ, Wagner M. "The Subsequent Artificial Neural Network (SANN) Approach Might Bring More Classificatory Power to ANN -Based DNA Microarray Analyses." Bioinformatics 2004; 20- 18, p. 3544 -3552. .

[9] Furey TS et al. "Support Vector Machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 2000; 16, p.906 – 914.

[10] NCBI, "The Sequence Read Archive," 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/sra.

[11] NCBI SRA Toolkit, [Online]. Available: https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/

[12] Boruta Feature Selection in R (Article) [Online]. Available: https://www.datacamp.com/community/tutorials/feature-selection-R-boruta.

[13] Kluytmans J, van Belkum A, Verbrugh H (July 1997). "Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks". Clinical Microbiology Reviews. 10 (3): 505–20. doi:10.1128/CMR.10.3.505. PMC 172932. PMID 9227864.

[14] Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG (July 2015). "Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management". Clinical Microbiology Reviews. 28 (3)