2022 IEEE 22nd International Conference on Nanotechnology (NANO)
July 4-8, 2022
Palma de Mallorca, Spain

Classification of DNA Sequences: Performance Evaluation of Multiple Machine Learning Methods

Yiren Wang[1], Vikram Khandelwal[2], Arindam K. Das[3], M.P. Anantram[4]

***Abstract*— Polymerase chain reaction (PCR) has long been the mainstay in genetic sequencing and identification. Irrespective of whether short read or long read technologies are adopted, PCR methods are generally time consuming and expensive. Recently, an all-electronic approach, the so-called Single Molecule Break Junction (SMBJ) method, has been proposed as a possible alternative to PCR. In this article, we evaluate the performance of four different classifier models on the current signatures of ten short strand sequences, including a pair that differs by a single mismatch. We find that a gradient boosted tree classifier model achieves impressive accuracies, ranging from approximately 96% for molecules differing by a single mismatch to 99.5% otherwise.**

I. INTRODUCTION

Identification of single molecules based on the value of their conductance is important for applications such as biomarker identification, disease detection, and chemical sensing. Conductance spectra through single molecules are however extremely noisy because of the stochastic and complex interactions between the substrate, sample, environment, and the measuring system. A large standard deviation in conductance values can obfuscate the ability to identify different molecules. In this talk, we discuss our recent effort in using machine learning methods to identify DNA strands based on the measured conductance [1] and subsequent effort in moving to smaller sample sizes.

Current traces are obtained from Single Molecule Break Junction (SMBJ) [1] measurements. Figure 1(a) shows a conceptual schematic of the SMBJ experimental setup. The experiment commences with the scanning microscope tip probing the conducting substrate, *hopefully* making contact with the DNA strand on the substrate. The tip is then gradually pulled away from the substrate. Given an appropriate voltage bias, the current between the tip and substrate is recorded as a function of time. Figures 1(b) and (c) show representative current traces with and without DNA binding between the tip and the substrate. Without any molecular binding, the current trace exhibits a predominantly exponential decay, as illustrated in Figure 1(b). Deviations from this behavior, as illustrated in Figure 1(c), are generally indicative of a successful molecular binding and a 'valid' experiment. Our hypothesis is that unique signatures of the DNA molecule exist in the conductance (ratio of current to applied voltage) traces, specifically, conductance probability distributions (histograms), and automated classification of DNA strands should be possible using statistical and/or machine learning (ML) based approaches.

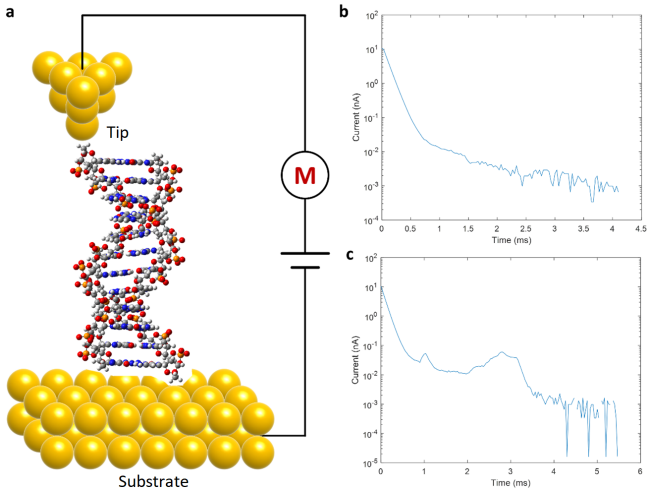


Fig. 1: (a) Schematic of the single-molecule break junction (SMBJ) experimental approach. (b) Sample current trace with no DNA binding between the gold electrode and the substrate. (c) Sample current trace with DNA binding. Figure from [2].

II. DATA AND CLASSIFICATION METHODOLOGIES

We use ten datasets, numbered S1 to S10, of experimentally obtained current traces to evaluate our classifiers. Each dataset contains a mix of valid (with molecular bonding) and invalid (without molecular bonding) traces. Some salient features of the datasets are as follows. First, although S2, S6, S7, S8, and S9 are of the same strand, current measurements were recorded using three different bias voltages, 0.01 V, 0.10 V, and 0.20 V. It is known that a variation in the applied bias induces a lateral shift in the conductance distributions. Second, S4 and S5 are mismatches of S3, which corresponds to mRNA from E.coli:O157:H7 with its fully matched DNA duplex and is known to produce both Shiga toxins (Stx) 1 and 2. S4 and S5 have the same DNA complement as S3. The mRNA from S4 corresponds to E.coli:O175:H28, has a single mismatch at base 14 of S3 (A is substituted by G), and is known to lead to Stx 2. The mRNA from S5 corresponds to E.coli:E1a, has a single mismatch at base 8 of S3 (C is substituted by T) and is nontoxic as it does not produce either

[1]Dept. of Electrical and Computer Engineering, University of Washington, Seattle, USA. ethanwyr@uw.edu
[2]Interlake High School, Bellevue, USA. vikram.a.khandelwal@gmail.com
[3]Dept. of Computer Science and Electrical Engineering, Eastern Washington University, Cheney, USA. adas@ewu.edu
[4]Dept. of Electrical and Computer Engineering, University of Washington, Seattle, USA. anantmp@uw.edu

2022 IEEE 22nd International Conference on Nanotechnology (NANO) | 978-1-6654-5225-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/NANO54668.2022.9928773

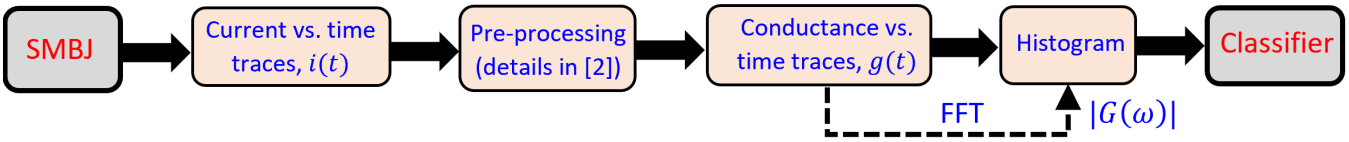


Fig. 2: Flow diagram of our approach for DNA sequence classification.

Stx 1 or 2. Complete details of the ten datasets can be found in [2].

Our current experimental data was generated from B-form DNA molecules with an average height of 4.08 nm, not including the carbon linkers at the two ends. All sequences in our current data are 12 bp or 15 bp long. It is known that AT rich sequences and those with a longer length tend to exhibit lower conductivity. The impact of these three parameters on classifier performance will be addressed in a subsequent paper.

We trained our classifier models using two different target class labeling schemes. In the first scheme (TLS-1), unique DNA strands were assigned different class labels, irrespective of the voltage bias used for current measurement during SMBJ experiments, resulting in six target classes. In the second scheme (TLS-2), the datasets are assigned unique class labels based on the (strand, voltage bias) tuple, resulting in eight classes. Additional details and justification for using the different labeling schemes are available in the Supplementary section of [2].

Figure 2 illustrates our approach for DNA sequence identification. Some pre-processing was necessary to filter out the invalid traces and reduce the noise in the raw current traces. Details of the pre-processing stage can be found in [2]. The histograms for each target class can be constructed either from the set of conductance vs. time traces, $\{g(t)\}$, or the set of corresponding conductance magnitude spectra, $\{|G(\omega)|\}$, where $|G(\omega)|$ is the magnitude of the Fourier Transform of $g(t)$. *Henceforth, we will refer to* $|G(\omega)|$ *simply as the conductance spectrum*. We evaluate the performance of four different classifier models: leftmargin=*

- *Approach*-1: Extreme gradient boosting (XGboost) on 600-bin histograms constructed from H time traces sampled randomly from the set $\{g(t)\}$.
- *Approach*-2: XGboost on 600-bin histograms constructed from H conductance spectra sampled randomly from the set $\{|G(\omega)|\}$.
- *Approach*-3: Multilayer perceptron (MLP with two hidden layers, $600-64-8(6)-8(6)$ with ReLu activation; numbers within parentheses refer to TLS-1) on 600-bin histograms constructed from H time traces sampled randomly from the set $\{g(t)\}$. From our experience, shallower networks tended to perform better than deeper networks, possibly due to the inherently noisy nature of the conductance traces.
- *Approach*-4: A distance based classification scheme on 600-bin histograms constructed from H time traces sampled randomly from the set $\{g(t)\}$. For each target class, one '*large sample template histogram*' is constructed using *all* training data for that class. A test histogram (constructed from H samples), say t, is assigned to class i if the Euclidean distance between t and H_i is the smallest, where H_i is the large sample template histogram of the i^{th} class.

As we will see shortly, the parameter H plays a critical role in classifier accuracy. This is because, for relatively large values of H (say 30), the inherent noise in individual conductance traces (or in the conductance spectra) is smoothed considerably, leading to more accurate histograms and enhanced classifier accuracy. Figure 3 shows the how the 'quality' of conductance distributions can differ depending on the value of H. From a practical perspective, achieving high accuracies with small H is highly desirable in order to minimize the time and cost associated with data collection.

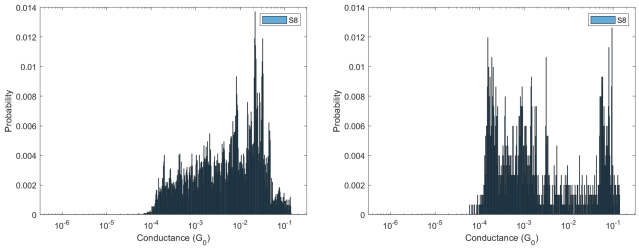
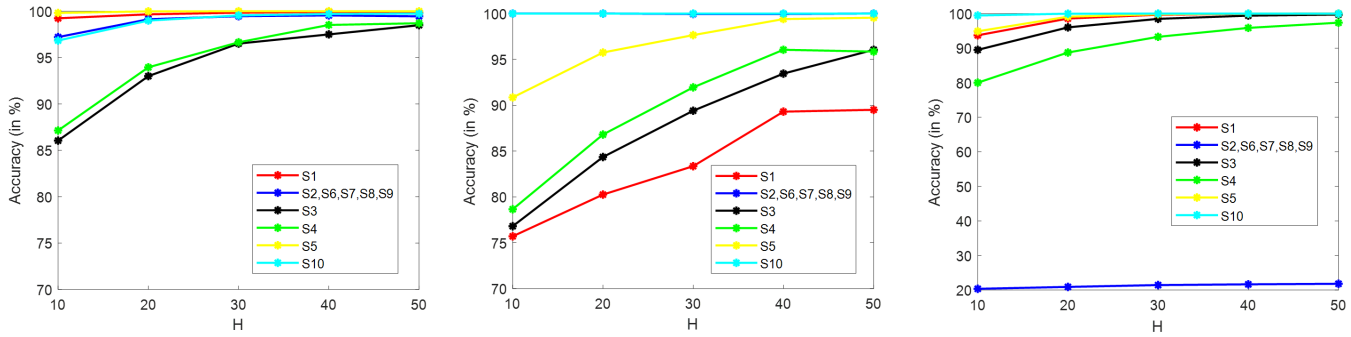


Fig. 3: Representative conductance histograms: (Left) $H = 30$, (Right) $H = 10$.

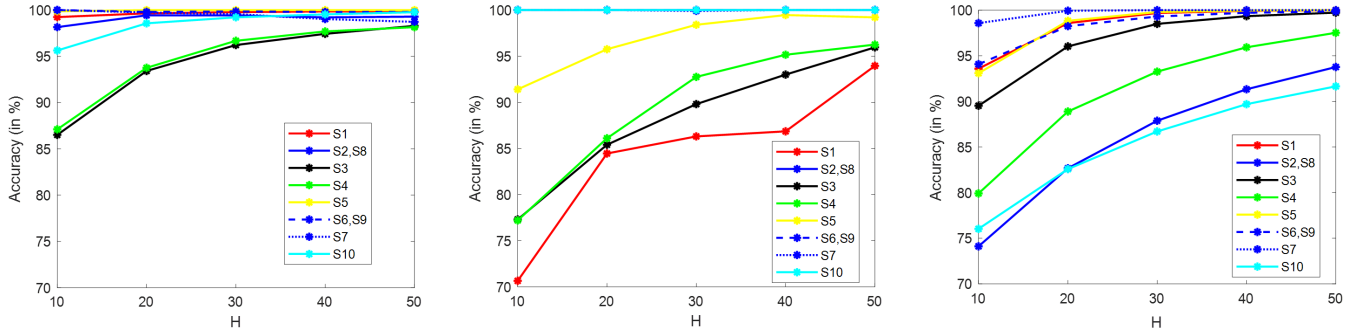
XGboost [3] is a fast and scalable implementation of a gradient boosted decision tree framework [4]. Gradient boosting is an ensemble learning method wherein weak base learners (usually decision trees) are added sequentially, one at each iteration, to minimize a suitably defined loss function evaluated on the previous learner. For details on gradient boosting, we refer the reader to [5]. We used the Python implementation of the XGboost package [6], with $N_{est} = 200$ and $D_{est} = 2$, where N_{est} denotes the number of trees/estimators and D_{est} denotes the depth of each tree/estimator.

III. Simulation Results

Figure 4 shows the overall accuracies of the classifiers (excluding the MLP approach) as a function of H. First, we observe that the accuracy of all four approaches drops sharply for lower values of H. Second, Approach-1 is the most accurate and offers $> 95\%$ accuracy for all classes for both labeling schemes. Third, for the finer 8-class labeling scheme (TLS-2), the simple distance based method (Approach-4) is only slightly outperformed by XGboost operating on conductance spectra (Approach-2). However, for



(a) Plot of overall classifier accuracy vs. H for target labeling scheme TLS-1: (Left) Approach−1, (Middle) Approach−2, (Right) Approach−4. In the middle panel, the plots for S2/S6/S7/S8/S9 (blue) and S10 (cyan) are virtually indistinguishable.



(b) Plot of overall classifier accuracy vs. H for target labeling scheme TLS-2: (Left) Approach−1, (Middle) Approach−2, (Right) Approach−4. In the middle panel, the plots for S2/S6/S7/S8/S9 (blue) and S10 (cyan) are virtually indistinguishable.

Fig. 4: Plot of overall accuracy vs. H for three classifiers with two different target labeling schemes.

the coarser 6-class labeling scheme (TLS-1), Approach−4 performs poorly for class S10, irrespective of the value of H, although the accuracies of the other five classes are surprisingly good.

Figure 5 shows the detailed confusion matrices for all four classifier models and both labeling schemes, when $H = 30$. Interestingly, the MLP model (Approach−3) struggles to distinguish between S4 and S5, which are both mismatches of S3. Another major source of confusion for the MLP, as well as the distance method (Approach−4), is between the strand corresponding to datasets S2/S6/S7/S8/S9 and the strand corresponding to S10. In our simulations, we have observed that boosted trees consistently outperform multilayer perceptrons. We conjecture that this is possibly due to noisy characteristics of the data. From Figure 5, we also observe that XGboost operating with histograms constructed from the conductance spectra (Approach−2) perform reasonably well, except for S1 (which is the molecule 'octanedithiol' and not a DNA/RNA strand) vs. S5. A possible reason why Approach−2 doesn't work quite as well as Approach−1 is that, since we are using only the magnitude spectrum, the information content in $\{|G(\omega)|\}$ is less than the information content in $\{g(t)\}$.

IV. Conclusion

In this paper, we have demonstrated that DNA molecules can be classified extremely accurately using ML methods operating on experimental quantum transport data. Typical classification accuracies for molecules which are structurally different exceed 99.9%. Even in the case of DNA-RNA hybrids with a single base pair mismatch, our best method is able to differentiate between the classes with an overall accuracy of over 96%. Our analysis demonstrates the potential of combining current spectra and ML methods as a diagnostic tool for real-time detection and classification of genetic sequences.

References

[1] Li, Y., Artés, J.M., Demir, B. *et al.* Detection and identification of genetic material via single-molecule conductance. *Nature Nanotech* **13**, 1167–73 (2018). https://doi.org/10.1038/s41565-018-0285-x

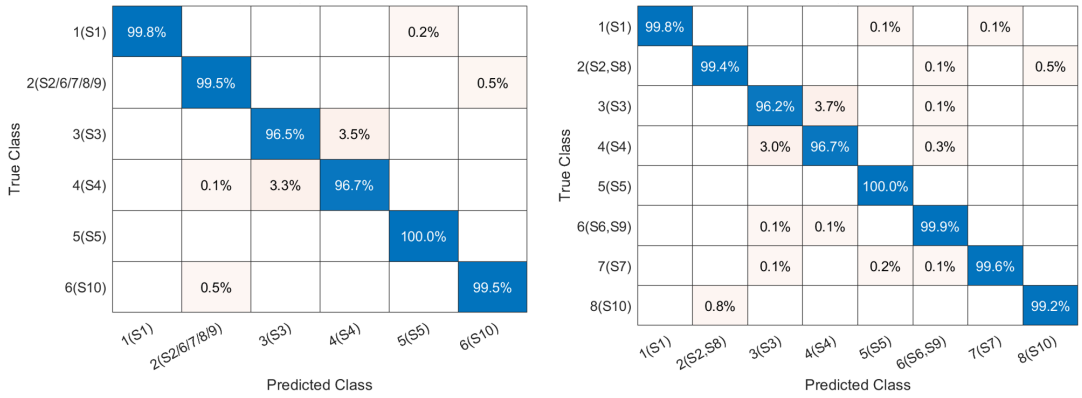
[2] Wang, Y., Alangari, M., Hihath, J. *et al.* A machine learning approach for accurate and real-time DNA sequence identification. *BMC Genomics* **22**, 525 (2021). https://doi.org/10.1186/s12864-021-07841-6

[3] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 42, pp. 785–94, 2016. doi:10.1145/2939672.2939785

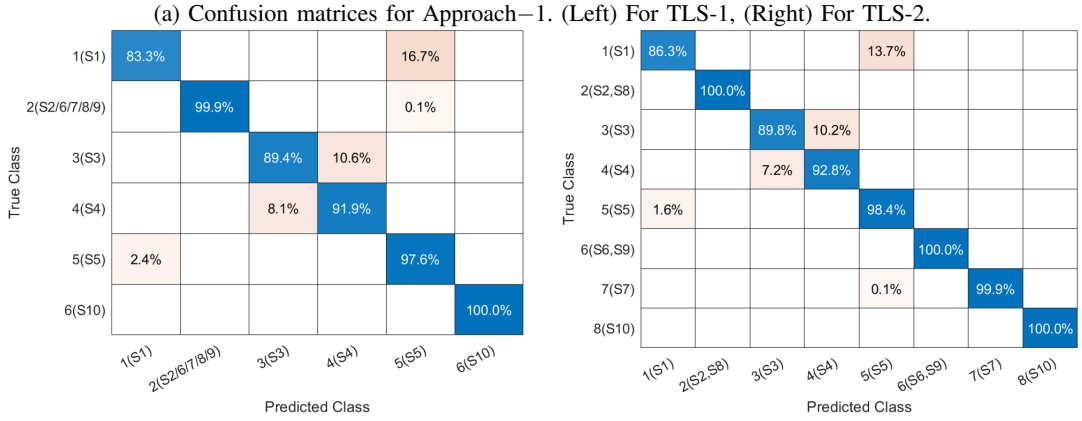
[4] Friedman J. H. Greedy function approximation: A gradient boosting machine. *Ann Stat*, 29:1189–232, 2001. doi:10.1214/aos/1013203451

[5] Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer New York; 2001. doi:10.1007/978-0-387-21606-5

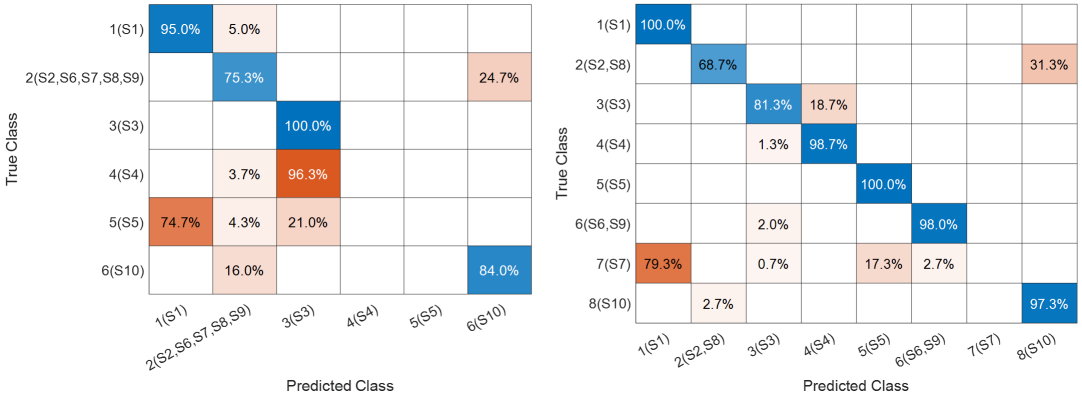
[6] XGboost Python Package — xgboost 1.3.0-SNAPSHOT documentation n.d. `https://xgboost.readthedocs.io/en/latest/python/index.html` (accessed September 18, 2020)



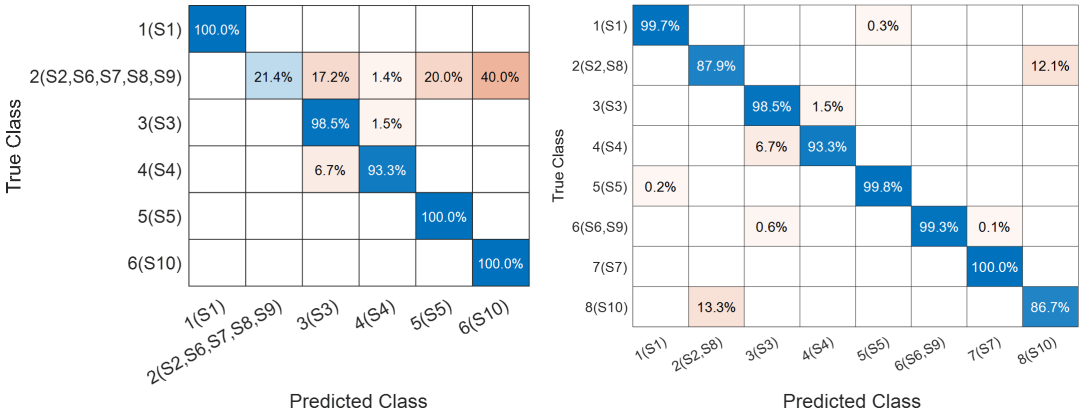
(a) Confusion matrices for Approach−1. (Left) For TLS-1, (Right) For TLS-2.



(b) Confusion matrices for Approach−2. (Left) For TLS-1, (Right) For TLS-2.



(c) Confusion matrices for Approach−3. (Left) For TLS-1, (Right) For TLS-2.



(d) Confusion matrices for Approach−4. (Left) For TLS-1, (Right) For TLS-2.

Fig. 5: Confusion matrices for four different classifiers with two different target labeling schemes, $H = 30$.