A DEEP LEARNING MODEL FOR PREDICTING NEXT GENERATION SEQUENCING DEPTH FROM DNA SEQUENCE

Dr.P.VijayaBharati[#1], K.LeelaKumari[#2], M.Dharani[#3], L.Poojitha[#4], P.SudhaSri[#5]

Department of Computer Science & Engineering, Vignan's Institute of Engineering For Women, Duvvada, Visakhapatnam.

ABSTRACT

The main method for genomics and molecular diagnostics, as well as more recently as readout for DNA information storage, is targeted high-throughput DNA sequencing. The variable hybridization kinetics of the oligo nucleotide probes employed to enrich gene loci of interest lead to non-uniform coverage, which raises the cost and lowers the sensitivity of sequencing. Here presenting a deep learning model (DLM)for estimating DNA probe sequence-based Next-Generation Sequencing (NGS) depth. The DNA nucleotide identities and the estimated likelihood of the nucleotide being unpaired are both inputs to DLM's bidirectional recurrent neural network. A 39,145-plex panel for human single nucleotide polymorphisms(SNP),a2000-plexpanelforhumanlongnon-codingRNA(lncRNA),anda7373-plexpaneltargetingnonhumansequencesforDNAinformation storage.

Keywords: Deep Learning, Next Generation Sequencing, Single-plex hybridization, Strand displacement rate constants, DNA Sequencing.

INTRODUCTION

The haploid human genome contains more over 3 billion DNA nucleotides, making thorough sequencing of the complete genome uneconomical for clinical use. Instead, targeted sequencing—in which a collection of DNA hybridization probes is created to bind and enrich the DNA areas of interest—is commonly used by researchers and diagnostic labs. However, the binding kinetics and thermodynamics of each DNA oligo nucleotide probe in a targeted sequencing panel often differs depending on the target. As a result, the enrichment efficiencies for various genetic loci will be drastically different when a panel of DNA probes is haphazardly designed and produced. The sequencing depth of a locus, which measures the number of NGS reads that contain it, directly relates to the sensitivity of NGS to that locus. The sensitivity at low-depth loci is either decreased by nonuniform sequencing depth or requires additional sequencing to ensure that every locus is read to the necessary minimum depth. Although labour-intensive and time-consuming, empirical tuning of an NGS panel's probe sequences and concentrations is now necessary. While creating a computer technique for calculating sequencing depth based on probe sequence, it is crucial to take into account both DNA sequence data and biophysical features. Model performance may suffer if the biophysical characteristics of DNA are ignored. What a fantastic summation! In fact, the area of bioinformatics has made substantial use of machine learning to glean useful information from voluminous biological data. Complex biological data can contain patterns and relationships that are hard or impossible for people to manually find. Machine learning techniques can be utilised to do this. Additionally, they can be utilised to create predictive models that assist researchers in deciphering the workings of biological systems and forecasting future events. The data must, however, appropriately reflect the biology being investigated because machine learning models are only as effective as the data they are

trained on. To ensure that machine learning models generalise effectively to new data, thorough validation and testing are necessary because they can be susceptible to overfitting.

BASICKNOWLEDGEOFDNA

An organism's whole DNA inventory is contained in its genome. All living things have genomes, yet their sizes vary widely. For instance, the human genome is divided into 23 chromosomes,which is similar to organizing an encyclopaedia into 23 volumes. Further more, each human genome contains more than 6 billion" base pairs, "or individual characters, if you count all of them. It is a substantial compilation.Over 6 billion characters or letters makeup a human genome. If you imagine the genome(the entire DNA sequence) as a book, it contains around 6 billion "A,""C,""G," and "T" letters. Now, each person has a distinct genome.None the less,the majority of the human genomes share similar features, according to scientists. Genomic science is a data-driven field that heavily relies on machine learning toidentify patterns in data and derive new biological theories. Yet, more potent machine learning models are needed to be able to draw new conclusions from the volume of genomics data that is growing exponentially. Deep learning has effectively rebuilt industries like computer vision and natural language processing by utilizing massive datasets.It has evolved into the method of choice for a variety of genomics modeling applications,including determining how genetic variation affects gene regulation processes like DNA receptivity and splicing. One of the most widely used technologies in the life sciences is gene sequencing. The sequencing platform with the best through put and lowest price is currently HiSeqXTen.The development of the sequencing industry has been considerably aided bythe introduction of technology and its commercialization. Sequencing is becoming more and more prevalent due to the technology's rapid advancement and ongoing cost reduction. The foundation of sequence data mining is sequence similarity, and this is an area of research where sequence similarity bioinformatics is quite useful. The degree of resemblance between sequences is referred to as sequence similarity.It is thought that two sequences may have homology if there is more than 30% similarity between them. The structures and functions of the homologous sequences may be comparable due to the fact that they share a common evolutionary ancestor.

SEQUENCING TECHNOLOGY

Sequencing technology has gone through three stages of development along with biological information technology. The first generation sequencing technology refers to both the chain termination approach proposed by Sanger and the chain degradation method proposed by Gibert. Sanger sequencing is still frequently employed in current conventional sequencing applications and verification, but its true large-scale application is greatly hampered by its extremely high sequencing cost and low throughput. After more than 40 years of technological growth, sequencing technology has achieved remarkable progress. Figure1 illustrates how sequencing technology has advanced. The second generation of sequencing technology, represented by the 454 technology, was born in 2005 as a result of ongoing research efforts. These sequencing devices have the capacity to simultaneously evaluate billions of sequencing reactions. The second-generation sequencing technique is a form of linked sequencing that considerably increases sequencing speed while substantially lowering sequencing cost.Second-generation sequencing technology is currently the dominant force in the market for scientific research, according to Watson (2014). It has a low price, hence it is extensively utilized. Oxford single molecule sequencing technology and PacBio's SMRT technology,which are examples of third generation sequencing technology,wereintroducedin2011.The

Most note worthy aspect of third-generation sequencing technology is single molecule sequencing. For use on a broad scale, this technology must be updated and changed regularly.Personalized medicine is undergoing a revolution thanks to sequencing technology, which offers high through put possibilities.

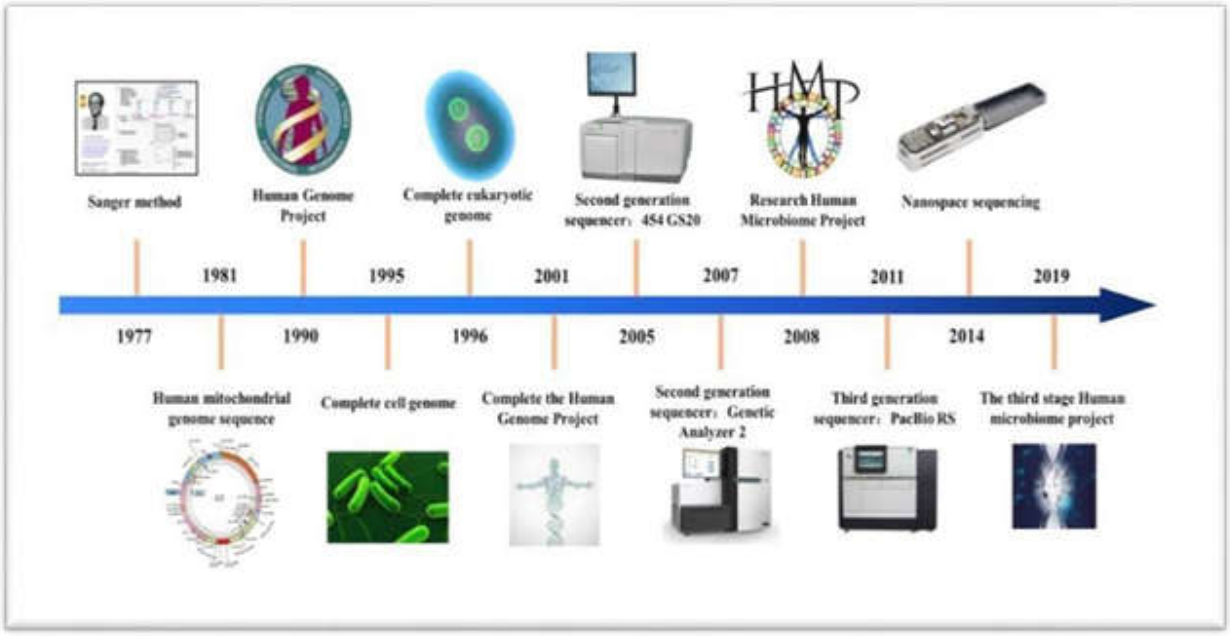


Figure1:History of sequencing technology

DATA STRUCTURE OF DNA SEQUENCE

Studies on biological sequences have demonstrated that they are not random, unordered strings. They are made up of a series of interconnected smaller components. There are four different types of deoxyribo nucleotides that link the DNA sequence (bases). The base order of DNA molecules affects their diversity. Figure 1.1 depicts the double helix structure of DNA. Only certain bases on the opposing strand of the DNA double helix can form bonds with nitrogen-containing bases on one strand. The fundamental building block of a DNA sequence is a base pair, which is what this process is commonly referred to as. DNA sequence data differ from other types of data in several ways, chiefly including:

Non-numeric characters (A, T, C, and G) make up the DNA sequence data.

The length of various sequences varies significantly.

DNA sequence data contains its unique biological significance.

Before performing any data analysis, the appropriate data pre treatment must be carried out becausethere were some sequencing errors and noise in the sequence data. Some sequences have only a few dozen characters, while others are very long and up to hundreds of megabytes.

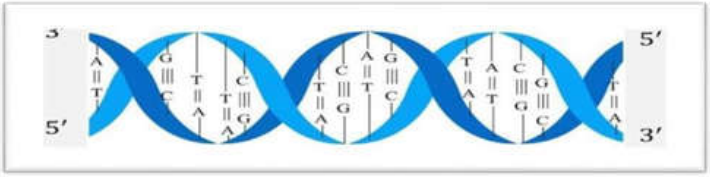


Figure2:Double helix of DNA

DNA SEQUENCE CODING

In order to create a matrix input model training, it is necessary to translate the string sequence used to represent the DNA sequence into a numerical value. Sequential encoding, one hot encoding, and k-mer encoding are the three main techniques for sequence encoding (Choong and Lee, 2017). Table 1 displaysthe traits of the three DNA encoding techniques.

Table1:Common ways of encoding DNA sequences

Encoding method	Features
Sequential encoding	This method encodes each base as a number. For example, change [A,T,G,C] to [0.25, 0.5, 0.75, 1.0], and any other character can be recorded as zero.
One-hot encoding	This method is widely used in deep learning methods. For example, [A,T,G,C] will become [0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0,0,0]. These coded vectors can be connected or turned into a two-dimensional array.
K-mer encoding	First take a longer biological sequence and decompose it into k-length overlapping fragments. For example, if we use a segment of length 6, "ATGCATGCA" will become: "ATGCAT," "TGCATG," "GCATGC," "CATGCA."

Sequential encoding performs similarly to one-hot encoding, but requires substantially less training time. Deep learning techniques frequently employ one-hot encoding, which is ideal for algorithms like CNN (convolutional neural networks). Moreover, one-hot encoding performs well across a variety of data sets, but high performance necessitates an appropriate CNN. In some evaluation data sets, matrices that reflect ordinal codes perform the best. The effectiveness of CNN in identifying DNA patterns depends on the accurate encoding and representation of the sequence. The successful performance of the ordinal coding approach demonstrates that the single-point coding method still has space for development. It is thought that two sequences have a homologous relationship if there is more than 30% similarity between them. As a result, if the two sequences are quite similar, they probably share an evolutionary ancestor. The function of the unknown sequence can also be predicted if a sequence that is similar to it can be located among sequences with known functions.

Neglecting our considerable understanding of DNA biophysics and relying solely on DNA sequence data to simulate sequencing depth versus probe sequences would most certainly result in subpar model performance. Although such expert systems are typically labour - intensive to create and display low generalizability to neighbouring problems, we also wish to avoid lengthy feature building and curation. Thus, we chose to strike a middle ground where we used just a few local (individual nucleotide-level) and global (oligonucleotide molecule- level) properties that can be fully autonomously computed by the well-known DNA folding software Nupack. Here, we developed a deep learning model (DLM) for predicting the sequencing depth of three NGS panels: one with 39,145 probes against human single nucleotide polymorphisms (abbreviated as SNP panel), one with 2000 probes against human long non- coding RNA (abbreviated as lncRNA panel), and one with 7373 probes against synthetic sequences created artificially

for infancy (abbreviated as synthetic panel). Although its probes were individually created and experimentally tested using the same library preparation approach as the SNP panel, the lncRNA panel serves as an independent test set for the SNP panel. The DLM is built on a recurrent

DEEP LEARNING MODEL

Deep learning can be seen as a subset of machine learning. It is a discipline where independent learning and development depend on the study of computer algorithms. Artificial neural networks, which are used in deep learning, are created to resemble how people think and learn. Recurrent Neural Networks are a Deep Learning method for modelling sequential data (RNN). Before to the development of attention models, RNNs were the go-to recommendation for handling sequential data. A deep feedforward model can need particular parameters for each component of the sequence. Moreover, it might not generalise to sequences of varying length. The behaviour of RNNs, which are built from feed forward networks, is comparable to that of human brains. In conventional neural networks, all of the inputs and outputs are independent of one another. Nevertheless, there are times when prior words are required, such as when predicting the next word of a sentence, and it is therefore important to remember the prior words. RNN was developed as a result, and it used a Hidden Layer to solve the issue. The Hidden state, which retains specific information about a sequence, is the most crucial part ofan RNN. RNNs have a memory where they keep track of all the calculations' data. Since it generates the same result by carrying out the same operation on all inputs or hidden layers, it uses the same parameters for each input. The DLM is built on a recurrent neural network (RNN) architecture to more effectively capture short- and long-range interactions that may affect the efficiency and speed of capture inside the DNA probe sequence. Targeted high-throughput DNA sequencing is increasingly being used in cancer treatment and has taken over as the primary tool for biological and biomedical research.

Targeted sequencing has lately been investigated as a technique for random-access readout of information densely and permanently stored in DNA. The primary goal is to perform DLM training and validation on two types of datasets: DNA interaction kinetics rate constants and NGS sequencing depth. The measured single-plex kinetic rate constants for DNA strand displacement and hybridization can both be accurately predicted by the same model. The complexity and heterogeneity of the primary diseases that kill millions of patients each year throughout the world are revealed by rigorous genetic investigations in the postgenomic era. Innovation is the primary value driver for reversing this trend and enhancing health. The enormous complexity of the most prevalent and serious complicated diseases, such as cancer, is revealed by systematic investigations for entire human genome sequencing, which are made possible by current NGS-based systems. NGS platforms create a logical method towards the completion of a mutations catalogue for common diseases with enhanced data analysis quality and fast declining costs. The evaluation of the genetic changes underlying severe diseases like cancer forms the basis for the creation of genetic testing based on aetiology.

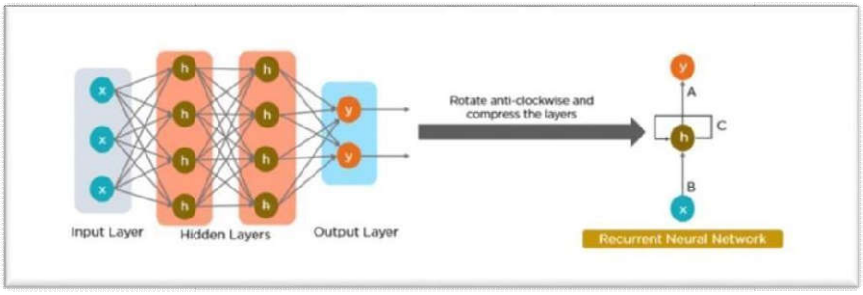


Figure3:Recurrent neural method work

PROBLEM STATEMENT

Even though the cost of DNA sequencing is falling rapidly over time, poor sequencing uniformity wouldwaste a significant portion of reads by sequencing high-depth targets twice and giving insufficient information on low-depth targets. Thus, there is a critical requirement to logically design NGS panels with great uniformity. Yet, the scale of NGS datasets makes the issue a good candidate for machine learning. Targeted high-throughput DNA sequencing is increasingly being used in cancer treatment and has taken over as the primary tool for biological and biomedical research. Targeted sequencing has lately been investigated as a technique for random-access readout of information densely and permanently stored in DNA. Large datasets are used by deep learning to automatically identify weakly correlated properties between inputs and outputs. As a result, deep learning has taken the lead in computer vision and other fields with enormous datasets at its disposal.

PROPOSED SYSTEM

The Sanger method is not comparable to the NGS technologies' massively parallel analysis, high throughput, and lower cost. Nonetheless, genomic sequences are useful thanks to NGS. The cost has decreased as the procedure has sped up (sequencing a human genome now takes only days to weeks). Large amounts of DNA can now be sequenced thanks to next-generation sequencing, such as all the parts of a person's DNA that contain the instructions needed to make proteins. DNA sequencing, a method for screening for genetic abnormalities, has enhanced the study of genetics by establishing the sequence of DNA building blocks (nucleotides) in a person's genetic codeWith the completion of the human genome project, 454 was introduced by 454 in 2005, Genome Analyzer was made available by Solexa the following year, and (Sequencing by Oligo Ligation Detection) SOLID was offered by Agen Court. Which are the three most common massively parallel sequencing technologies used in next- generation sequencing (NGS), all of which outperformed Sanger sequencing in terms of throughput, accuracy, and cost.

The DLM was independently trained using the SNP panel and the synthetic panel, and sequencing depths were predicted for each panel separately using cross validation. While the SNP panel and the lncRNA panel both employed the same library preparation technique, they were tested separately. Because each NGS library preparation method involves a wide range of unique experimental variables (experimental work-flow, sample type, hybridization

temperatures, etc.), we thought that these variables were outside the DLM's purview. Practically speaking, we anticipate that rather than aiming to increase consistency across different NGS library preparation methods, the majority of users will want to specifically optimize probe sequence and concentration. Nevertheless, we attempted training and forecasting using both the SNP panel and the synthetic panel, and Supplementary Note 3 provides a summary of the outcomes. In Supplemental Data 1-3, probe sequences and measured read depth are shown. Two components make up each of our NGS datasets: characteristics derived from probe sequences and read depth as determined by a single NGS library. For the SNP panel and the synthetic panel, we divided the data into 20 classes at random, and as shown in Fig.5.1.a, predictions for each class (representing 5% of the whole dataset) were derived using a DLM trained on the remaining 19 classes (representing 95% of the total dataset). In order to assess prediction accuracy, the 20-fold cross validation predictions employed a total of 20 DLMs. The global features and the log10(Depth) inside each training set were standardized to have a mean of 0 and a standard deviation of 1. The global features of each training set's matching validation set were normalized using its mean and standard deviation, and the model predictions were rescaled to reflect their original mean and standard deviation. The DLM contains over 300,000 weight parameters. In order to address the vanishing gradient issue for deep NNs16, these were present via Xavier initialization (uniformly distributed weights with standard deviation dependent on the number of parameters in a layer). In order to update the network weights during training, we iteratively minimized the Loss using gradient descent and an Adam optimizer17. This loss is a function of the mean squared difference between the experimental and expected log sequencing depths. After each hidden layer of the FFNN, we added an additional dropout layer to reduce overfitting. In this layer, 20% of the parameters are randomly chosen and aren't allowed to update duringeach training cycle. Tensorflow18 was used to create the DLM, and its hyper- parameters include batch size (999), learning rate (0.0001), node dropout fraction (20%), GRU hidden nodes (128), and FFNN hidden nodes (256 and 128). The settings indicated above seem to produce the quickest training time and the highest prediction performance out of the roughly 50 sets of varied hyper-parameter values we tested. The SNP panel and the synthetic panel both have training epochs of 250 and 1000, respectively. For the SNP panel, feature generation using unpack requires about 0.5 seconds per probe sequence on a standard desktop computer, and training time for each epoch is roughly 10 seconds while using less than 3 gigabytes of memory from a graphics processing unit.

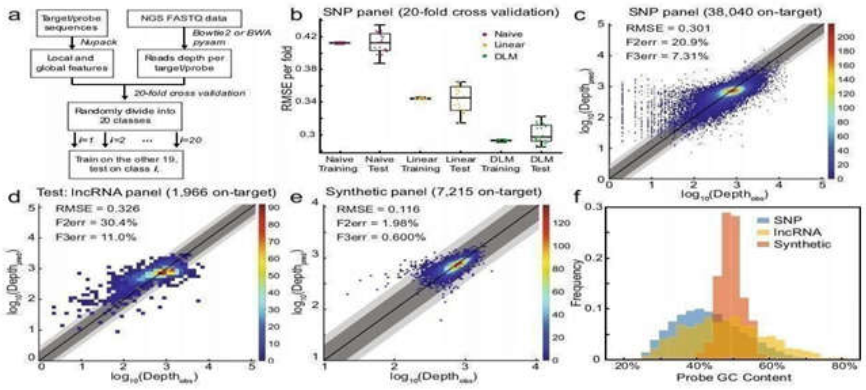


Figure 4: Cross-validation and independent test results of the DLM on predicting NGS depth.

Our knowledge of DNA and NGS leads us to conclude that the yield and speed of DNA probe hybridization, as well as chemistry-specific biases, are the main determinants of NGS read depth. As aresult, our DLM ought to be useful in estimating the DNA hybridization rate constants. We subsequently used the DLM to the prediction of a related DNA mechanism, strand displacement, in order to push our DLM even more and to demonstrate how successful our DLM methodology is20,21. We use time-based fluorescence data, which allows us to monitor a single target and probe species with great time and yield resolution, as opposed to NGS studies, in which hundreds of DNA probes and targets are simultaneously hybridizing. In Supplemental Data 4 and 5, targets, blockers, and probe sequences are listed.

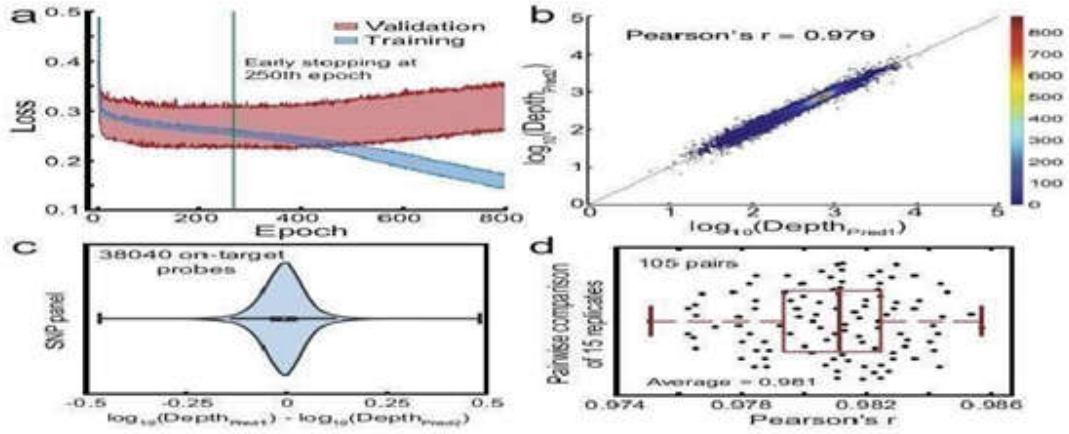


Figure 5: Reproducibility of DLM predictions for SNP panel

To determine which features actually make a difference in prediction accuracy, we examined various DLM that were deficient in those traits. The applications where a reduced DLM model performs noticeably worse than our default DLM are highlighted in red. The columns with blue highlights show the condensed DLMs that perform almost as well as our default DLM.

NGS depth (Human)	0.301	0.310	0.301	0.310	0.315	0.302	0.304	0.309	0.301
NGS depth (Synthetic)	0.12	0.15	0.12	0.16	0.12	0.12	0.12	0.11	0.12
Hyb. rate constant	0.32	0.47	0.45	0.30	0.38	0.32	0.36	0.32	0.32
Displacement rate constant	0.43	0.67	0.77	0.40	0.52	0.40	0.52	0.42	0.42

RMSE

Pr Local Feature - Base pair open probability $p_{exposed}$
S Local Feature - Sequence Identity(A/T/C/G)
E_T Global Feature - standard free energy of Target $\Delta G^0(T)$
E_P Global Feature - standard free energy of Capture Probe $\Delta G^0(P)$
E_{TP} Global Feature - standard free energy of TP complex $\Delta G^0(TP)$
T Global Feature - Reaction Temperature

Feature Used
Feature NOT Used

Figure 6: Assessing the importance of different components of the DLM to prediction accuracy, measured by RMSE.

Although certain characteristics are superfluous for the particular applications we are considering, we see that the majority of features enhance performance in at least one application. where there is a lack of carefully selected data for particular issues. The design of narrowly focused prediction software is guided by expert knowledge in expert system machine learning approaches based on considerable human feature construction and curation. Nevertheless, these systems typically perform poorly when applied to identical challenges. In order to avoid the trap of labor-intensive and problem-specific model and design, we limited the inputs to a small number of global and local features that can be automatically computed based on DNA sequence. Considering the thermodynamics of DNA model. We think that there is a good chance that the base-pair probability values predicted by Nupack have significant error because of inaccuracy or incompleteness and the fact that we could not realistically take into account the intermolecular interactions from all 3+ billion nucleotides in the human genome. Future models that anticipate base-pair accessibility in a highly complex and heterogeneous solution more precisely may be essential to enhancing the DLM's prediction accuracy. Since many years ago, the two basic mechanisms serving interaction between DNA sequences— hybridization and strand displacement— have been understood to be present in all living things as well as DNA-based biotechnology platforms like PCR and microarray. On the other hand, numerous data bases have been built since Next- Generation Sequencing technology emerged. Nowadays, people can access many areas of interest at once and get thousands of times more data.

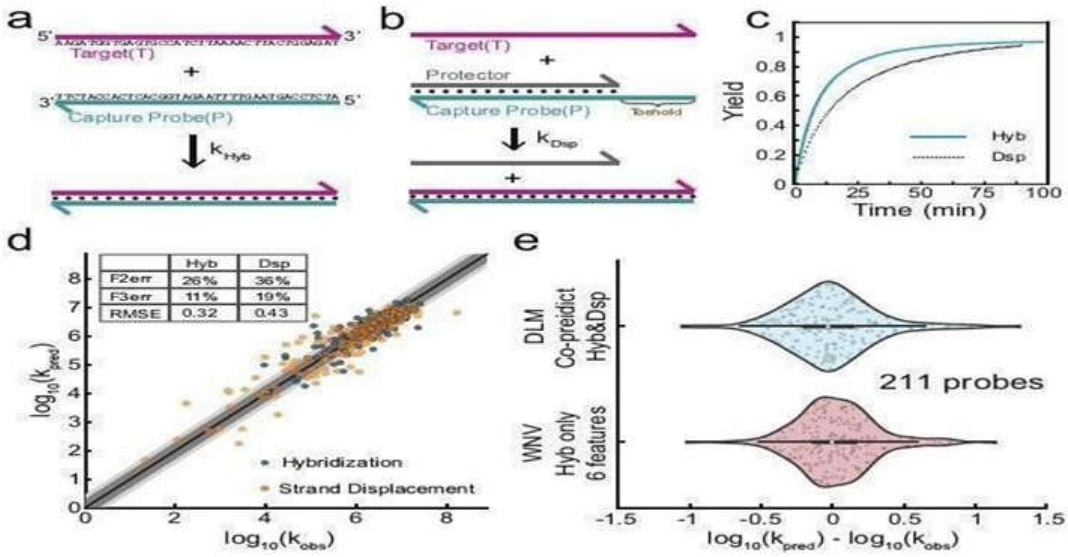


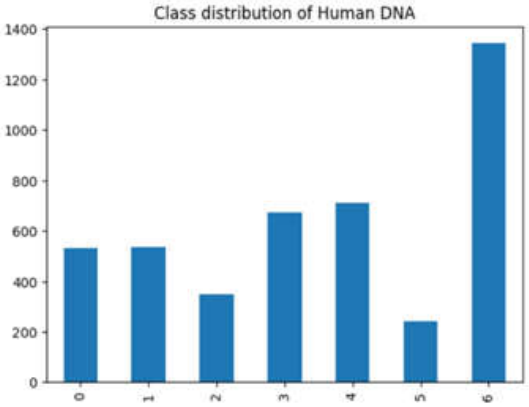
Figure7:Applying DLM to the prediction of single-plex hybridization and strand displacement rate constants

LITERATURE SURVEY

- Target-enrichment strategies for next-generation sequencing.
- The thermodynamics of DNA structural motifs
- Predicting DNA hybridization kinetics from sequence
- Neural network based multi-classifier system for gene identification in DNA sequence
- Gene structure prediction using informationon homologous protein sequence

RESULT

Class distribution of Human DNA

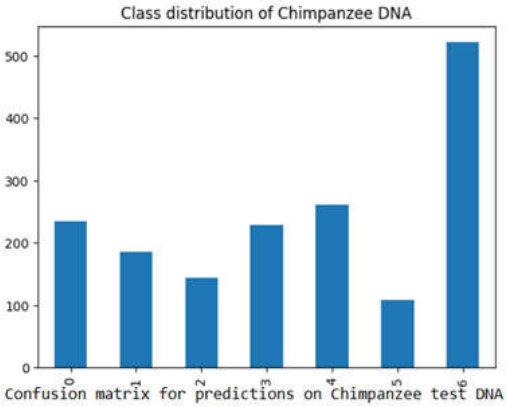


Confusion matrix for predictions on human test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	99	0	0	0	1	0	2
1	0	104	0	0	0	0	2
2	0	0	78	0	0	0	0
3	0	0	0	124	0	0	1
4	1	0	0	0	143	0	5
5	0	0	0	0	0	51	0
6	1	0	0	1	0	0	263

accuracy = 98
precision = 98
recall = 98
f1 = 98

- **Class distribution of Chimpanzee DNA**

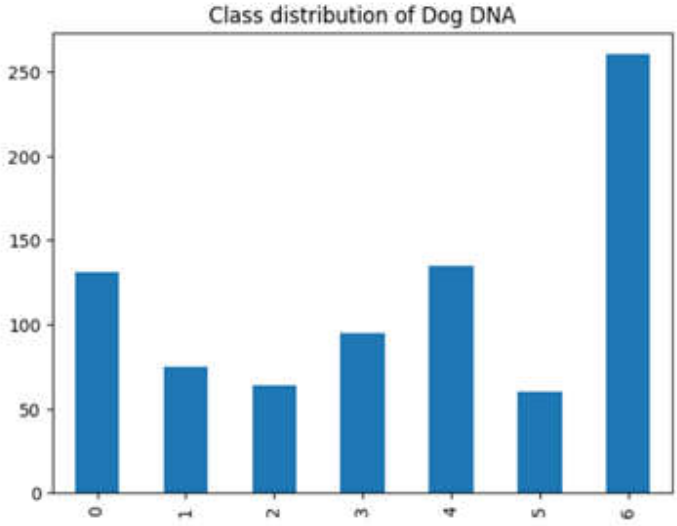


Confusion matrix for predictions on Chimpanzee test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	232	0	0	0	0	0	2
1	0	184	0	0	0	0	1
2	0	0	144	0	0	0	0
3	0	0	0	227	0	0	1
4	2	0	0	0	254	0	5
5	0	0	0	0	0	109	0
6	0	0	0	0	0	0	521

accuracy = 99
precision = 99
recall = 99
f1 = 99

- **Class distribution of Dog DNA**



Confusion matrix for predictions on Dog test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	127	0	0	0	0	0	4
1	0	63	0	0	1	0	11
2	0	0	49	0	1	0	14
3	1	0	0	81	2	0	11
4	4	0	0	1	126	0	4
5	4	0	0	0	1	53	2
6	0	0	0	0	0	0	260

accuracy = 92
precision = 93
recall = 92
f1 = 92

CONCLUSION AND FUTURE SCOPE

After analyzing the previous studies, it is obvious to us that DNA hybridization and NGS sequencing technologies are quite significant. We also recognize the need to improve the DLM's prediction precision. With DLM, we avoided the trap of labor-intensive and problem- specific model creation by limiting our inputs to a small collection of global and local features that can be automatically computed based on DNA sequence. The measured single-plex kinetic rate constants for DNA strand displacement and hybridization can both be accurately predicted by the same model. Unknown species and complex diseases can be understood thanks to NGS technologies. Despite the fact that many businesses use various platforms, each with its own unique benefits. The objectives will be to improve assembly sequencing precision, shorten processing times, and improve analysis algorithm efficiency.

REFERENCES

1. Cheng, W. Y., Chen, H. & Morrison, J. Kinetics of DNA replication in a dicentric X chromosome formed by long arm to long arm fusion. Human Genet. 56, 71–79 (1980).

2. Reynaldo, L. P., Vologodskii, A. V., Neri, B. P. & Lyamichev, V. I. The kinetics of oligonucleotidereplacements. J. Mol. Biol. 297, 511–520 (2000).

3. Zhang, D. Y. & Winfree, E. Control of DNA strand displacement kinetic using toehold exchange. J.Am. Chem. Soc. 131, 17303–17314 (2009).

4. Zhang, J. X. et al. Predicting DNA hybridization kinetics from sequence. Nat.Chem. 10, 91–98 (2018).

5. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. J.Comput.Chem. 32, 170–173 (2011)

6. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.(2014).

7. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015).

8. Pinto, A., Chen, S. X. & Zhang, D. Y. Simultaneous and stoichiometric purification of hundreds of oligonucleotides. Nat. Commun. 9, 1–9 (2018).

9. Simmel, F. C., Yurke, B. & Singh, H. R. Principles and applications of nucleic acid strand displacement reactions. Chem. Rev. 119, 6326–6369 (2019).

10. Zhang, D. Y. Towards domain-based sequence design for DNA strand displacement reactions. In International Workshop on DNA-Based Computers, pp. 162–175 (Springer, Berlin, Heidelberg, June 2010).

11. Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M. & Nguyen, M. A practical approach to RT- qPCR–publishing data that conform to the MIQE guidelines. Methods 50, S1– S5 (2010).

12. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR analysis: realtime monitoring ofDNA amplification reactions. Nat. Biotechnol. 11, 1026 (1993).

13. Das, J. et al. An electrochemical clamp assay for direct, rapid analysis of circulating nucleic acids in serum. Nat. Chem. 7, 569 (2015