Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An interpretable machine learning method for detecting novel pathogens

Xiaoyong Zhao (microhello@gmail.com)
Beijing Information Science and Technology University https://orcid.org/0000-0003-0345-3331
Ningning Wang
Beijing Information Science and Technology University

Research article

Keywords: interpretable, machine learning, bacterial pathogen

Posted Date: March 17th, 2020

DOI: https://doi.org/10.21203/rs.2.20477/v2

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: According to the World Health Organization (WHO), infectious diseases continue to one of the leading causes of death worldwide. Since the core microbiota flora of humans is largely diverse and horizontal gene transfer (HGT), it is very challenging to determine whether a particular bacterial strain is commensal or pathogenic to humans. With the latest advances in next-generation sequencing (NGS) technology, bioinformatics tools and techniques using NGS data have increasingly been used for the diagnosis and monitoring of infectious diseases. Even if the biological background is not available, the machine learning method can still infer the pathogenic phenotype from the NGS readings, independent of the database of known organisms, and being studied intensively.However, previous methods have not considered opportunistic pathogenic and interpretability of black box model, are not well suited for clinical requirements.

Results :In this study, we proposed a novel interpretable machine learning approach (IMLA) to identify the pathogenicity of bacterial genomes: human pathogens (HP), opportunistic pathogenicity (OHP) or non-pathogenicity(NHP), then use the following model-agnostic interpretation methods to interpret model: feature importance, accumulated local effects and Shapley values, due to the model interpretability is essential for healthcare applications. To our knowledge, our paper is the first attempt to infer opportunistic pathogenicity and explain the model.

Conclusions: According to the simulation results, our approach IMLA can be a great addition to detect novel pathogens.

1 Background

1.1Motivation

According to the World Health Organization (WHO), infectious diseases continue to one of the leading causes of death worldwide. Recent research has shown that humans are heavily colonized by thousands of microbes that are harmful, harmless, or beneficial to human health. (Qin et al., 2010; Hooper and Gordon, 2001). Since human core microbiota is very diverse, it is particularly challenging to determine whether a particular bacterial strain is pathogenic to humans.

Currently, the gold standard for determining infectious agents is the Koch hypothesis established in the 19th century, which requires isolation and cultivation of microbial strains (Sassetti et al., 2003; Young et al., 1984). However, the culture period is long and many pathogens are difficult to culture, so it is difficult to meet the clinical requirements.

Another phylogeny-based method classifying bacteria into human pathogenicity was to look for some molecular features(Falkow,1997), two-component systems (Stock et al.,2000) and secretion systems(Hacker and Kaper.,2000) . However, there are two problems with these methods. One is that these features are exchanged between pathogenic and avirulent strains of the same or different species due to horizontal gene transfer (HGT) (Frost et al., 2005; Frost et al., 2008; HoSui et al. , 2009). Another is that there are some virulence genes, although they cannot directly determine virulence, but they are essential for bacterial response in the host to survive and evade the body's immune system(Wassenaar and Gaastra,2001; Paine et al., 2002).

With the latest advances in next-generation sequencing (NGS) technology, DNA sequencing has become the state-of-the-art in pathogen detection (Lecuit and Eloit, 2014; Calistri and Palù, 2015), The amount of data in bacterial sequence database is rapidly accumulating. (Benson et al., 2015; O'Leary et al., 2016), bioinformatics tools and techniques using NGS data have increasingly been used for the diagnosis and monitoring of infectious diseases detection (Lecuit and Eloit, 2014; Calistri and Palù, 2015). Even if the biological background is not available, the machine learning method can still infer the pathogenic phenotype from the NGS readings, independent of the database of known organisms, and being studied intensively.

1.2Related Work

Existing methods for predicting pathogenicity can be broadly divided into two types, genome-based (Miller et al., 2013; Byrd et al., 2014; Naccache et al., 2014; Deneke et al., 2017; Jakub M. Bartoszewicz et al. , 2019) and protein-based (Iraola et al., 2012; Cosentino et al., 2013; Eran et al.,2018), as described below. The method we propose in this article belong to the former category.

Genome-based methods use the raw genome reads as input. Miller et al. developed several tools for mapping NGS reads to reference genomes and for classification (Miller et al, 2013; Byrd et al, 2014; Naccache et al, 2014). However, these tools make the taxonomy rather than phenotype prediction, and severely affected the basic data set covering taxonomy, and cannot be used to predict new pathogens.

Recently, some genome-based phenotype prediction methods were published.

Deneke et al. proposed a random forest-based pathogenicity prediction method PaPrBaG (Deneke et al., 2017). It predicts the pathogenicity of novel unknown bacterial pathogens by training a large number of labelled pathogens and non-pathogenic bacteria. Compared to other methods, PaPrBaG can be predicted based on NGS data with very low genomic coverage, while other methods are based on the similarity of the known reference genomes.

Another recent method for predicting pathogenicity is Deep Learning Approach to Pathogenicity Classification (DeePaC), includes a universal, extensible framework for neural architectures ensuring identical predictions for any given DNA sequence and its reverse-complement (Jakub M. Bartoszewicz et al. , 2019). It combines the reverse-complementary architecture with the integration of predictions for both mates in a read pair results, and designs a reverse-complementary convolutional neural network and Long Short-Term Memory (LSTM), reducing the error rate by nearly half compared to the latest technology.

Although these new methods are more accurate, the application of any of these methods to mission-critical contexts remains problematic due to un-interpretability of black-box model and neglect of opportunistic pathogenicity.

Protein-based methods characterize the phenotype of the microbe by the presence or absence of members of the protein families (PFs) in its genomes, provided that the assembled genomes are available.

Iraola et al. proposed the first large-scale application of the protein-based method BacFier, training a Support Vector Machine (SVM) model to predict bacterial virulence based known families of orthologous genes (Iraola et al., 2012). This method depends on a virulence factor database that annotates virulence at the genetic level and is therefore limited to specific proteins known to be associated with virulence, ignoring many unannotated genes whose sequences are available and possibly virulence (or anti-virulence) function related.

Other protein-based pathogenicity prediction tools that create and annotate PF based on their frequency of occurrence in pathogenic or non-pathogenic organisms, without depending on pre-established databases (Cosentino et al., 2013; Eran et al.,2018).

Cosentino et al. (Cosentino et al., 2013) developed a web server PathogenFinder for predicting bacterial pathogenicity using proteomics, genomes or raw reads (https://cge.cbs.dtu.dk/services/PathogenFinder/) . The pathogenicity of bacteria depends on the proteome known to be involved in pathogenicity.This web server utilizes a selection of proteins created without annotated function or known involvement in pathogenicity. It can predict pathogenicity for all taxonomic groups of bacteria with 88.6% accuracy. The approach of the program is not biased with known pathogenicity. Therefore the program could be used to discover novel pathogenicity factors. However, the step of clustering proteins into PFs step in this method has computational bottleneck.

Eran et al. proposed a machine learning method BacPaCS for bacterial pathogenicity classification through sparse support vector machine (sparse-SVM). By fully automating the training of clinically relevant data, the calculation time is greatly shortened, and the training data set is much larger than before (Eran et al. 2018). Experimental results in a clinically relevant data set containing only human host bacteria showed that BacPaCS showed high accuracy in distinguishing between pathogenic bacteria and non-pathogenic bacteria.

However, the human body has plenty of long-term coexistence of microbes, these microbes in many cases do not exhibit pathogenicity, but under certain conditions, will be pathogenic to humans. Previous methods have not considered opportunistic pathogenic and they are not well suited for clinical requirements.

In this paper we propose a novel interpretable machine learning approach IMLA for classifying unidentified bacterial genomes as human pathogens, opportunistic pathogenicity or none of them, then use the following model-agnostic interpretation methods to interpret model: feature importance, accumulated local effects and Shapley values, as model interpretability is essential for healthcare applications.

2 Methods

We will describe the construction of the data set in Section 2.1 and describe the classification method in Sections 2.2 and 2.3.

2.1. Dataset

2.1.1. Extracting data

To the best of our knowledge, there are currently no publicly available standard data on human pathogenic microorganisms.We extracted metadata from the Pathosystems Resource Integration Center (PATRIC) and the Integrated Microbial Genomes (IMG) . Both databases provide researchers with a variety of metadata on microbial genome projects.

We downloaded the IMG and PATRIC database on 05/07/2019, IMG contained ~71 thousand sequenced bacterial genomes, PATRIC contained ~220 thousand sequenced bacterial genomes.

We further identified the human colonization and whole genome sequences (WGS) data by finding ‘Humans sapiens’ or ‘Homo sapiens’ in the ‘host name’ column and finding ‘WGS’ in the ‘Genome Status’ column, then merged and deduplicated both databases by NCBI Taxon ID column, and finally downloaded fasta format files of sequences according to taxon ids from NCBI website.

2.1.2. Labelling method

To infer the pathogen label, we created heuristic rules to label data based on metadata of the IMG and PATRIC databases, comparing with BacPaCS (Cosentino et al., 2013) and PaPrBaG (Deneke et al., 2017). The heuristic method was described as follows:

1. We labelled genomes as opportunity human pathogens (OHP) if they satisfied any of the following criteria: One of the fields ‘Isolation Source’, ‘Host Health’, or ‘Comments’ includes an OHP term, as defined in Table 1.
2. Excluding the generated OHP list, we labelled genomes as human pathogens (HP) if they satisfied any of the following criteria:

- The ‘Disease’ field is not empty and does not contain an OHP/Commensal term, as defined in Table 1.
- One of the fields ‘Isolation Source’, ‘Host Health’, or ‘Comments’ includes an HP term. In addition, the same fields cannot include any of the non-human pathogens (NHP) terms, as defined in Table 1.

3. Excluding the generated OHP and HP list, we labelled genomes as NHP if they satisfied any of the following criteria: One of the fields ‘Isolation Source’, ‘Host Health’, or ‘Comments’ includes an NHP term.

The term groups were used for the criteria, as defined in Table 1:

We labelled 87,574 genomes as human pathogens (HP), 107 genomes as opportunity human pathogens (OHP), 670 genomes as non-human pathogens (NHP) and 954,763 as inconclusive.

2.1.3. Balancing dataset

Since most bacterial samples are collected from clinically sick individuals, most of the sequencing bacteria are HP bacteria. In our data set, the ratio of HP:NHP:OHP bacteria is 200:6:1, and the data is imbalanced.

However, machine learning algorithms can be affected by imbalanced data set, and when the sample size ratio varies greatly, it may seriously affect the learning process of the algorithm, resulting in poor classification results.

These methods are often used to imbalanced learning: data-level methods (over-sampling, under-sampling and hybrid-sampling.) and algorithm-level methods (cost-sensitive learning, ensemble methods etc). Since the labeled training samples are difficult to obtain, and the generated synthetic samples have no definite meaning and are difficult to explain, so the data-level methods are not suitable, we choose cost-sensitive learning method to improve the importance of the minority class samples, set the weights of HP/NHP/OHP class to 1/200,1/6,1 respectively.

2.2 Training the Model

The workflow of our method is shown in Figure 1.

2.2.1 Extracting features

The effective features are very important for building a high performance classifier.

Bacterial genomes are densely packed with proteins (Patthy,1999), and since the protein sequences are more conserved evolutionarily than the DNA sequences, wherein the peptide can thus provide valuable information reliably.

We extracted protein physicochemical properties features and outer membrane proteins features to capture the information implied in sequencing reads.

- physicochemical properties

Amino acids are the basic unit of protein. The physicochemical properties of some amino acids have been determined by extensive experiments and theoretical studies.

AAindex is a numerical index database that assigns a fraction of each property (usually based on the peptide's secondary structure) to each residue, representing various physicochemical and biochemical properties of amino acid and amino acid pairs (Kawashima et al.,2008). The latest version 9.2 contained 566 properties when we accessed the website on 05/07/2019 (AAindex, 2019). AAindex is widely used to analyze the physicochemical and biochemical properties of protein sequences.

We selected the 32 of 566 properties as features of the model according to Deneke et al., 2017, as follows in Table 2.

The values of the feature were obtained by multiplying the amino acid frequency and its associated index score.

- outer membrane proteins features

Membrane proteins are proteins that are part of the interaction with biological membranes (Johnson and Cornell,1999; Alenghat and Golan,2013), that are the targets of over 50% of all modern medicinal drugs (Overington et al.,2006). It is estimated that 20–30% of all genes in most genomes encode membrane proteins (Krogh et al.,2001; Liszewski, 2015). Bacterial outer membrane vesicle (OMV)-mediated delivery of proteins to host cells is an important mechanism for host-pathogen communication (Koeppen et al.,2016). For example, the highly conserved Gram-negative outer membrane protein murein lipoprotein (MLP) has been detected in the plasma of septic patients (Hellman et al, 2000; Suzuki et al, 1978) and is the primary antigen that induces systemic infection. Both IgG in mice and humans have a homeostasis (Melody Y. Zeng et al., 2016) .Bertoletti and Gehring have shown that the removal of 26-30 AA from preS1, an outer membrane protein on the surface of HBV, cannot infect hepatocytes, confirming its significance in HBV infection (Bertoletti and Gehring,2009).

Protein secondary structure is the three-dimensional form of local segments of proteins, it’s simpler than tertiary structure. Membrane proteins can be predicted based on the protein secondary structure. Many methods have been developed to help predict protein secondary structure definitions and biochemical functions of proteins from its amino acid sequence (Cheng et al.,2005; Morten et al.,2012; Magnan and Baldi,2014; Mills et al.,2015;Yang et al.,2015;Zheng et al.,2019).

Read sequence was first translated into a peptide sequence, then we used standalone TMHMM version 2.0 (Krogh et al.,2001) to predict the transmembrane helices in proteins, and then used the normalized frequencies of transmembrane groups and helices as features of the model.

2.2.2 Generating Model

The Gradient boosting decision trees (GBDT), also known as Gradient Boosting Machine (GBM) is a boosting integrated learning model with the advantages of highly customizable flexibility, fast training and parallelization, easy to adjust and interpretable (Friedman,2001).

The GBDT methods have been widely adopted in academia, industry and competitive data science due to their most advanced performance in many machine learning tasks, being almost the gold standard for structured data sets.

Researchers have done some comparative study of the three most popular GBDT packages: XGBoost, LightGBM and Catboost.

XGBoost is an efficient, flexible and portable distributed gradient enhancement library. It implements machine learning algorithms under the Gradient Boosting framework and provides parallel tree enhancements to quickly and accurately solve many data science problems such as regression, classification, and ranking (Chen et al., 2016).

LightGBM is also a tree-based gradient boosting framework that uses a novel Gradient-based One-Side Sampling (GOSS) technique to filter out the data instances for finding a split value, while XGBoost uses the pre-sorting algorithms and histogram-based algorithms to calculate the optimal split (Ke et al.,2017).

Catboost is an open-source gradient boosting algorithm developed by Yandex team in 2017. It is a machine learning algorithm which allows users to handle categorical features for a large data set quickly and this differentiates it from XGBoost & LightGBM. Catboost can be used to solve regression, classification and ranking problems (Prokhorenkova et al.,2018).

The results show that Catboost outperformed the other two in terms of both speed and accuracy results for datasets with categorical features (Prokhorenkova et al.,2018; Anghel et al.,2018; Nguyen et al.,2018;Pulicherla et al.,2019). In addition, for datasets with a large number of features, XGBoost may not work due to memory limitations, and Catboost will converge into a good solution in the shortest time.

Therefore, this paper chooses to implement the IMLA model based on CatBoost, using the above features and pathogenic labels to train a classifier for each genome in the training data set.

2.2.3 Evaluation metrics

Different metrics are used to measure the accuracy of machine learning algorithms. Sensitivity and specificity are metrics which are well known for balanced datasets. Sensitivity (also known as recall or true positive rate TPR) is the proportion of well-classified positive samples, while specificity (also known as true negative rate TNR) is the proportion of well-classified negative samples.

However, the above both metrics would result in misleading scores due to the imbalance of the data. We choose the Precision-Recall curve (PR AUC) and Matthews correlation coefficient (MCC) metrics to deal with imbalanced datasets:

Precision-Recall curve is the area under the precision-recall curve, is a more reliable and informative indicator than the receiver operating characteristic curve, especially for imbalanced datasets (Hand and Till, 2001; Saito and Rehmsmeier,2015).

MCC was a method introduced by biochemist Brian W. Matthews in 1975 to measure the quality of binary classification, it was more informative than other confusion matrix measures (such as F1 score and accuracy) , it has also been generalized to the multiclass case (Gorodkin,2004; Chicco,2017).

2.3 Interpreting Model

Doshi et al. define the interpretability of a machine learning system as an ability to explain to humans or present in understandable terms (Doshi et al., 2017). Interpretability is critical for model debugging, detection bias, interpersonal collaboration, compliance, and mission-critical applications such as healthcare, where understanding, verification, editing, and trust models are important (Caruana et al., 2015). .

There are three methods for solving this problem: interpretable model, model-specific interpretation method and model-agnostic interpretation method (Molnar et al.,2019). Interpretable models often have a big disadvantage: predictive performance is reduced compared to other machine learning models. The downside of the model-specific interpretation method is that it also binds you to a model type and makes it difficult to switch to other models. Model-agnostic interpretation methods have great flexibility because they separate interpretation from machine learning models, and machine learning developers are free to use any machine learning model when applying interpretation methods to any model (Ribeiro et al., 2016).

We use the following model-agnostic interpretation methods to interpret model:

2.3.1 Features Importance

The displacement feature importance metric was introduced by Breiman for random forests (Breiman, 2001). Based on this idea, Fisher et al. proposed a model-agnostic version of feature importance and called it model dependency (Fisher et al., 2018).

Feature importance provides a highly compressed global insight into the behavior of the model, automatically considering all interactions with other features. When the value of the feature is modified, the prediction error of the model increases, which destroys the relationship between the feature and the true outcome.

2.3.2 Accumulated local effects

Accumulated local effects (D.W. Apley. 2016) describe how the feature affects the prediction of the machine learning model on average, even if the features are correlated, it still works. ALE plots are faster to compute and scale with O(n), and are a faster, more unbiased alternative to partial dependence plots (Friedman, 2001). The ALE plots are centered at zero, which makes the value of each point of the ALE curve the difference from the mean prediction, so the interpretability is very good.

The interpretation of the ALE diagram is clear. Given a value, the relative impact of the changed feature on the prediction can be derived from the ALE plots.

2.3.3 Shapley value

The Shapley value is a method in coalitional game theory that allocates payout to players based on their contribution to total payout, which is the average marginal contribution of the value of the feature in all possible coalitions (Shapley, 1953).

The Shapley value is the only interpretation based on solid theory (efficiency, symmetry, dummy and additivity axioms), and the difference between the predicted value and the average predicted value is fairly distributed between the feature values of the instance. Furthermore, the Shapley value allows for comparative interpretation without comparing the prediction to the average prediction of the entire data set, but rather comparing the Shapley value to a subset or even a single data point.

3 Results

Cross-validation strategy is a model validation techniques used to evaluate how the results of a predictive model will be generalized to an independent data set, estimating the performance of the predictive model in practice (Kohavi, 1995). We split dataset into train and test subsets with 7:3 ratio, both have same HP/NHP/OHP ratio, then evaluate the model using standard 10-fold cross-validation within the training set.

In the following, we furthered compared the performance of our method against the state-of-the-art in potential pathogenic prediction: PaPrBaG, BacPaCS and DeePaC. We trained PaPrBaG random forests, BacPaCS sparse-SVM and DeePaC RC-network using our dataset.

We used original authors settings for all methods(Deneke et al., 2017; Barash et al., 2018; Jakub M. Bartoszewicz et al. , 2019). Each classifier was trained on our x86_64 server with 12 Core CPU, 512GB RAM and Tesla V100 GPU, installed Ubuntu 16.04 and CUDA 10.0.

3.1 Implementation details

We implemented the architectures and all steps needed for data preprocessing, feature engineering, training, prediction and evaluation of the resulting models using python 3.6 with package pandas, NumPy, SciPy, scikit-learn, sci-imbalance, catboost, Tmhmm.py 1.2.2 for the transmembrane helix finder, as well as Biopython 1.72 for efficient fasta file preprocessing.

3.2 Hyperparameter tuning

Using a grid of parameter settings is currently the most widely used method for hyperparameter optimization, we used grid_search method provided by catboost, and set parameters grid: iterations=

[300,500,1000,1500], learning_rate=[0.01,0.03,0.05] and depth=[5,6,7,8,9,10]}, optimized hyperparameter result: iterations 1500,learning_rate 0.05 and depth 10.

3.3 Classifier Performance comparison

In order to compare with other classifiers, we merge OHP and NHP class into NHP, and thus transform it into binary classification problems, the weights of HP/NHP class set to 1/200,1 respectively for the balancing. The results of the evaluation as follows in Table 3.

The IMLA binary classifier outperforms all of the other binary classifiers in terms of TPR,PR AUC and MCC. The IMLA multiclass classifier's specificity is the best.

Thanks to the good performance of Catboost, the IMLA classifier training takes about 2 minutes,10-fold cross validation takes about 40 minutes, was the fastest method.

3.4 Model Interpreting

3.4.1 Features Importance

As follows Figure 2 plot the importance of the features.

The features importance scored by Shapley value, which is a general evaluation of feature relevance on the model as a whole. The chart considers the feature relevance for each prediction class. Also, for each feature, it shows how many features influence the fact aspects.

As an example, we can see that transmem_helix_ratio and transmem_group_ratio were by far the most influential features, which was consistent with what it was explained in Section 2.2.1.

3.4.2 Accumulated local effects

ALE plots for each feature are shown in Figure 3.

ALE plots describe how the feature affects the prediction of the machine learning model on average. The interpretation of ALE plots is clear, taking transmembrane group ratio as an example, it can be seen that when the transmembrane group ratio is greater than 0.2, the effect of this feature on the pathogenic prediction will remain unchanged.

3.4.3 Shapley value

There are some different Shapley diagrams to explain models from different aspects. Shapley summary plot is shown in Figure 4, Shapley model explainer plot is shown in Figure 5 and Shapley individual prediction explainer plot is shown in Figure 6.

Looking at transmem_group_ratio variable, we can see how lower ratio is associated with a big decrease in shap values. It is interesting to note that around the value 0.2-0.8, the curve starts to decrease again. It shows a perfect non-linear relationship.

Taking WERD780104, we can observe that shap values were almost 0 when the variable value was lower than 0.8, while on the other hand, the value one was associated mainly with a shap increase around 0.8.

This chart tries to explain the model as a whole. It essentially has all the samples plotted on the x-axis (in this case, ordered by similarity, but it can be changed in the combo box) and their prediction values plotted on the y-axis. Also, it has the individual contributions of each feature for each sample, based on feature value.

In this example, we selected sample number 9008 (x-axis), which has a prediction value of 1.363 (y-axis) and where transmem_helix_ratio and AURR980118 are the most relevant features — they push the prediction up though. However, by just hovering over other samples, it's possible to see how feature values and their impact change, as well as the predictions.

This chart can explain individual predictions. In this case, we selected the first sample of the test set. We can see that the model predicted a value of 0.18. Additionally, we can see which features contribute to getting that value higher (red) or lower (blue). In this case, QIAN880123 being equal to 0.6964 is the most defining feature that corroborates the target variable, meaning it pushes the prediction down. On the other hand, the values of transmem_helix_ratio and CHOP780207 improved the prediction value.

4 Discussion

In this paper, we explored the potential to predict the pathogenic phenotype of novel bacterial species directly from sequencing reads. We proposed a novel interpretable machine learning method for classifying novel bacterial genomes into human pathogens, opportunistic pathogenicity or non-pathogenicity, combining feature extraction with Catboost prediction, and then using the following model-agnostic interpretation methods to interpret the model: feature importance, accumulated local effects and Shapley values, given that model interpretability is essential for healthcare applications.

The training process of machine learning methods is highly dependent on the available pathogenic labels. Therefore, based on publicly available phenotypes and reliable pathogenicity information, we created a rule-based pathogenic annotation inference protocol and compiled a new bacterial genome dataset with labels.

5 Conclusions

In summary, the novel approach presented in this paper proposed a reliable and accurate classifier that quickly classifies HP, OHP and NHP. To our knowledge, this paper is the first attempt to predict opportunistic pathogenicity and to explain the model. Furthermore, how to measure the degree of interpretability, the retraceability and causability of method are very important in the healthcare domain(Holzinger et al.,2019; Holzinger et al.,2020), and these need to be further studied in the future.

Abbreviations

WHO: World Health Organization

HGT: Horizontal gene transfer

NGS: next-generation sequencing

IMLA: interpretable machine learning approach

HP: human pathogens

OHP: opportunity human pathogens

NHP: non-pathogenicity

PATRIC:Pathosystems Resource Integration Center

IMG: Integrated Microbial Genomes

DeePaC: Deep Learning Approach to Pathogenicity Classification

LSTM: Long Short-Term Memory

SVM: Support Vector Machine

WGS: whole genome sequences

GBDT: Gradient boosting decision trees

GBM: Gradient Boosting Machine

GOSS: Gradient-based One-Side Sampling

TPR: true positive rate

TNR: true negative rate

PRAUC: Precision-Recall curve

MCC: Matthews correlation coefficient

Declarations

Acknowledgements

The authors thank the anonymous referees for suggestions that helped improve the paper substantially.

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the National Natural Science Foundation of China (No. 61572079, No. 71701020) and the Beijing Municipal Education Commission of Science and Technology Plan Project (KM201711232018, KM201711232017).

Authors' contributions

ZXY conceived and developed the method. WNN analyzed the data. ZXY wrote the manuscript. ZXY and WNN reviewed and improved the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Qin,J. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, 59–65.
2. Hooper,L. V and Gordon,J.I. (2001) Commensal host- bacterial relationships in the gut. Science (80-.)., 292, 1115–1118.
3. Sassetti,C.M. et al. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol., 48, 77–84.
4. Young,R.A. et al. (1984) Genes for the major protein antigens of the leprosy parasite Mycobacterium leprae. Nature, 316, 450–452.
5. Falkow S (1997) What is a pathogen. ASM news 63: 359–365.
6. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annual review of biochemistry 69: 183–215.
7. Hacker J, Kaper JB (2000) Pathogenicity islands and the evolution of microbes. Annual Reviews in Microbiology 54: 641–679.

8. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology 3: 722–732.
9. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, et al. (2008) Variation in virulence among clades of escherichia coli O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences 105: 4868–4873.
10. HoSui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL (2009) The association of virulence factors with genomic islands. PLoS ONE 4: e8094.
11. Wassenaar TM, Gaastra W. (2001) Bacterial virulence: can we draw the line? FEMS Microbiology Letters 201: 1–7.
12. Paine K, Flower DR, et al. (2002) Bacterial bioinformatics: pathogenesis and the genome. Journal of molecular microbiology and biotechnology 4: 357–365.
13. Saeb, Amr T M.(2018) “Current Bioinformatics resources in combating infectious diseases.” Bioinformation vol. 14,1 31-35. , doi:10.6026/97320630014031
14. Benson,D.A. et al. (2015) GenBank. Nucleic Acids Res., 43, D30-5.
15. O’Leary,N.A. et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733–D745.
16. Byrd,A.L. et al. (2014) Clinical PathoScope: Rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics.
17. Naccache,S.N. et al. (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res.
18. Deneke,C. et al. (2017) PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Sci. Rep., 7, 39194.
19. Iraola,G. et al. (2012) Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. PLoS One, 7, e42144.
20. Cosentino,S. et al. (2013) PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS One, 8, e77302.
21. Eran Barash, et al.(2018) BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. Bioinformatics .
22. Bartoszewicz, Jakub M., et al. (2019) DeePaC: Predicting pathogenic potential of novel DNA with a universal framework for reverse-complement neural networks. BioRxiv : 535286.
23. Miller,R.R. et al. (2013) Metagenomics for pathogen detection in public health. Genome Med
24. Eschke, Kathrin; Trimpert, Jakob; Osterrieder, Nikolaus; Kunec, Dusan (2018). Mocarski, Edward (ed.). Attenuation of a very virulent Marek's disease herpesvirus (MDV) by codon pair bias deoptimization. PLOS Pathogens. 14 (1): e1006857. doi:10.1371/journal.ppat.1006857.
25. Mapleson, Daniel; Garcia Accinelli, Gonzalo; Kettleborough, George; Wright, Jonathan; Clavijo, Bernardo J. (2016). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics: btw663. doi:10.1093/bioinformatics/btw663.

26. Yakovchuk, P. (2006). "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix". Nucleic Acids Research. 34 (2): 564–574. doi:10.1093/nar/gkj454.

27. Bernardi, Giorgio (2000). "Isochores and the evolutionary genomics of vertebrates". Gene. 241 (1): 3–17. doi:10.1016/S0378-1119(99)00485-0.

28. Weber, Claudia C; Boussau, Bastien; Romiguier, Jonathan; Jarvis, Erich D; Ellegren, Hans (2014). "Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition". Genome Biology. 15 (12). doi:10.1186/s13059-014-0549-1.

29. Żemojtel, Tomasz; kiełbasa, Szymon M.; Arndt, Peter F.; Behrens, Sarah; Bourque, Guillaume; Vingron, Martin (2011). "CpG Deamination Creates Transcription Factor–Binding Sites with High Efficiency". Genome Biology and Evolution. 3: 1304–1311. doi:10.1093/gbe/evr107..

30. Karlin, Samuel (1998). "Global dinucleotide signatures and analysis of genomic heterogeneity". Current Opinion in Microbiology. 1 (5): 598–610. doi:10.1016/S1369-5274(98)80095-7.

31. Hershberg R, Petrov DA(2008). Selection on Codon Bias. Annual Review of Genetics. Annual Reviews; 42: 287–299. doi:10.1146/annurev.genet.42.110807.091442

32. Patthy, L(1999). Genome evolution and the evolution of exon-shu ing–a review. Gene 238, 103–114.

33. HE Bing, SONG Xiao-feng(2012). Progress in Prediction of Protein Ubiquitination Sites Based on theProtein Sequence. Progress in Modern Biomedicine

34. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M(2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res.36, D202-D205.

35. Cheng, A. Randall, M. Sweredoski, & P. Baldi(2005). SCRATCH: a Protein Structure and Structural Feature Prediction Server.Nucleic Acids Research, vol. 33 (web server issue), w72-76,.

36. Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu & Jinbo Xu(2012). Template-based protein structure modeling using the RaptorX web server. Nature Protocols 7, 1511–1522.

37. N. Magnan & P. Baldi (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity.Bioinformatics, vol 30 (18), 2592-2597.

38. L. Mills, P. J. Beuning, and M. J. Ondrechen(2015), “Biochemical functional predictions for protein structures of unknown or uncertain function,” Computational and Structural Biotechnology Journal, vol. 13, pp. 182–191.

39. J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang(2015). The I-TASSER Suite: Protein structure and function prediction. Nature Methods 12: 7-8

40. W Zheng, C Zhang, Q Wuyun, R Pearce, Y Li, Y Zhang(2019). LOMETS2: Improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. Nucleic Acids Research.

41. Johnson JE, Cornell RB (1999). "Amphitropic proteins: regulation by reversible membrane interactions (review)". Molecular Membrane Biology. 16 (3): 217–

35. doi:10.1080/096876899294544. PMID10503244.

42. Alenghat FJ, Golan DE (2013). "Membrane protein dynamics and functional implications in mammalian cells". Current Topics in Membranes. 72: 89–120. doi:10.1016/b978-0-12-417027-8.00003-9.

43. Overington JP, Al-Lazikani B, Hopkins AL (2006). "How many drug targets are there?". Nature Reviews. Drug Discovery. 5 (12): 993–6.

44. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes". Journal of Molecular Biology. 305 (3): 567–80. doi:1006/jmbi.2000.4315.

45. Liszewski K (2015). "Dissecting the Structure of Membrane Proteins". Genetic Engineering & Biotechnology News (paper). 35 (17): 1, 14, 16–17. doi:1089/gen.35.17.02.

46. Y. Zeng et al (2016)., “Gut Microbiota-Induced Immunoglobulin G Controls Systemic Infection by Symbiotic Bacteria and Pathogens,” Immunity, vol. 44, no. 3:647–658.

47. Koeppen K, Hampton T H, Jarek M, et al (2016). A novel mechanism of host-pathogen interaction through sRNA in bacterial outer membrane vesicles. PLoS pathogens, 12(6): e1005672.

48. Bertoletti A, Gehring A (2009). Therapeutic vaccination and novel strategies to treat chronic HBV infection . Expert Rev Gastroenterol Hepatol,3(5) : 561-569.

49. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol, 305(3):567–580.

50. Friedman J H (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics: 1189-1232.

51. Chen T, Guestrin C (2016). Xgboost: A scalable tree boosting system,Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM,: 785-794.

52. Ke G, Meng Q, Finley T, et al (2017). Lightgbm: A highly efficient gradient boosting decision tree,Advances in Neural Information Processing Systems.: 3146-3154.

53. Prokhorenkova L, Gusev G, Vorobev A, et al (2018). CatBoost: unbiased boosting with categorical features,Advances in Neural Information Processing Systems: 6638-6648.

54. Anghel A, Papandreou N, Parnell T, et al (2018). Benchmarking and Optimization of Gradient Boosted Decision Tree Algorithms. arXiv preprint arXiv:1809.04559.

55. Nguyen V K, Zhang W E, Sheng Q Z (2018). Identifying Price Index Classes for Electricity Consumers via Dynamic Gradient Boosting. International Conference on Web Information Systems Engineering. Springer, Cham: 472-486.

56. Pulicherla P, Kumar T, Abbaraju N, et al (2019). Job Shifting Prediction and Analysis Using Machine Learning,Journal of Physics: Conference Series. IOP Publishing, 1228(1): 012056.

57. Hand D J, Till R J (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning, 45(2): 171-186.

58. Saito T, Rehmsmeier M (2015). The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE.10(3):e0118432.

59. Gorodkin, Jan (2004). Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry 28 (5): 367–374.

60. Chicco D (2017). Ten quick tips for machine learning in computational biology. BioData Mining.,10 (35): 35.

61. Kohavi,R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Int. Jt. Conf. Artif. Intell., 14, 1137–1143.

62. Doshi-Velez F (2017), Kim B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint arXiv:1702.08608.

63. Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730.

64. Molnar, Christoph (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable,. https://christophm.github.io/interpretable-ml-book/.

65. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). Model-agnostic interpretability of machine learning. ICML Workshop on Human Interpretability in Machine Learning.

66. Breiman, Leo (2001).Random Forests. Machine Learning 45 (1). Springer: 5-32 .

67. Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2018). Model Class Reliance: Variable importance measures for any machine learning model class, from the ‘Rashomon’ perspective. http://arxiv.org/abs/1801.01489.

68. Apley, Daniel W (2016). Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468 .

69. Friedman, Jerome H (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics : 1189-1232.

70. Shapley, Lloyd S (1953). A value for n-person games. Contributions to the Theory of Games 2.28 : 307-317

71. M. Lundberg and S.-I. Lee (2017), A Unified Approach to Interpreting Model Predictions, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc, 4765–4774.

72. AAIndex (2019).AAindex: Amino acid index database. [ONLINE] Available at: https://www.genome.jp/aaindex/AAindex/list_of_indices. [Accessed 5 July 2019]

73. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Mueller, H (2019). Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9 (4), doi:10.1002/widm.1312.

74. Holzinger, A., Carrington, A. & Müller, H (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, 34 (2), doi:10.1007/s13218- 020-00636-z.

Tables

Table 1. The term group list

Group	Terms List (case insensitive)
HP terms	virulence, disease, superbug, patient, diarrhea, waterborne, foodborne, toxin, clinical, intensive, outbreak, infection, pathogen, water borne, food borne
NHP terms	healthy, probiotic, commensal, comparative, reference
OHP terms	opportunistic, condition, conditional, occasion, harmless
Commensal terms	healthy, Commensal, Periodontally healthy.

Table 2. The AAindex accession number and the description of features

AAindex accession number	description
AURR980102	Normalized positional residue frequency at helix termini N"' (Aurora-Rose, 1998)
AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose, 1998)
AURR980118	Normalized positional residue frequency at helix termini C" (Aurora-Rose, 1998)
BROC820101	Retention coefficient in TFA (Browne et al., 1982)
BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)
FAUJ880105	STERIMOL minimum width of the side chain (Fauchere et al., 1988)
FINA910103	Helix termination parameter at posision j-2,j-1,j (Finkelstein et al., 1991)
FUKS010109	Entire chain composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)
GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
GEIM800105	Beta-strand indices (Geisow-Roberts, 1980)
ISOY800107	Normalized relative frequency of double bend (Isogai et al., 1980)
KHAG800101	The Kerr-constant increments (Khanarian-Moore, 1980)
LEWP710101	Frequency of occurrence in beta-bends (Lewis et al., 1971)
MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga, 1976)
OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al., 1985)
PALJ810111	Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981)
PRAM820103	Correlation coefficient in regression analysis (Prabhakaran-Ponnuswamy, 1982)
QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)
QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)
QIAN880123	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)
QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
RACS770103	Side chain orientational preference (Rackovsky-Scheraga, 1977)
RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)
RICJ880101	Relative preference value at N" (Richardson-Richardson, 1988)
RICJ880117	Relative preference value at C" (Richardson-Richardson, 1988)
ROBB760107	Information measure for extended without H-bond (Robson-Suzuki, 1976)
SUYM030101	Linker propensity index (Suyama-Ohara, 2003)
TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)
TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga, 1977)
VASM830101	Relative population of conformational state A (Vasquez et al., 1983)
WERD780104	Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978)

Table 3. Performance results for all methods

Classifier	TPR	TNR	PR AUC	MCC	ACC
IMLA-Multiclass	0.91	**0.92**	0.91	0.87	0.92
IMLA-Binary	**0.99**	0.90	**0.99**	**0.89**	**0.94**
PaPrBaG	0.76	0.85	0.95	0.61	0.80
BacPaCS0.950.500.970.46DeePac	0.83	0.91	0.99	0.76	0.87
PaPrBaG	0.76	0.85	0.95	0.61	0.80

Figures

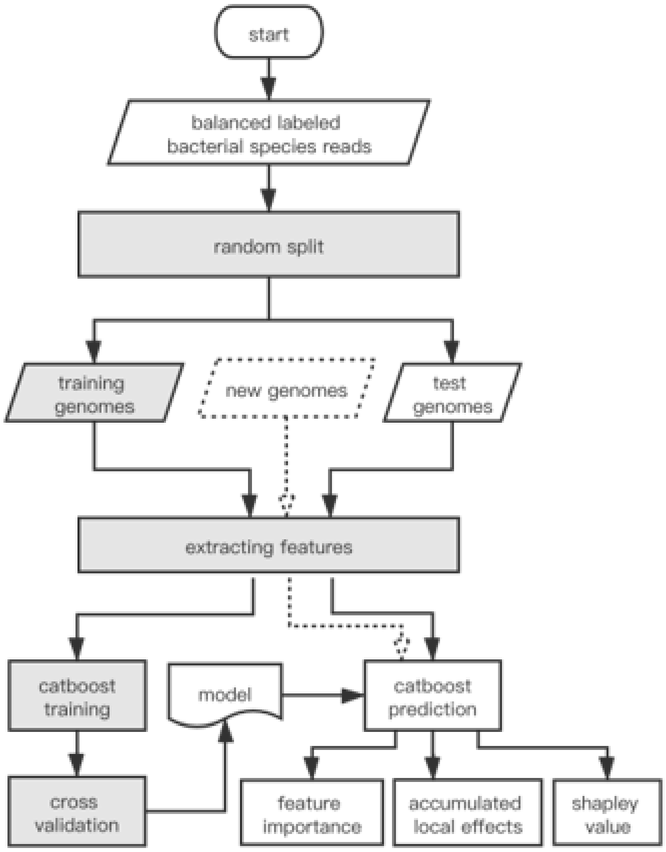

Figure 1

IMLA method workflow. Training and testing steps are outlined in continuous lines, and new genomes prediction steps are outlined in dashes. Input and output cells are colored white. Gray cells represent learning processes.

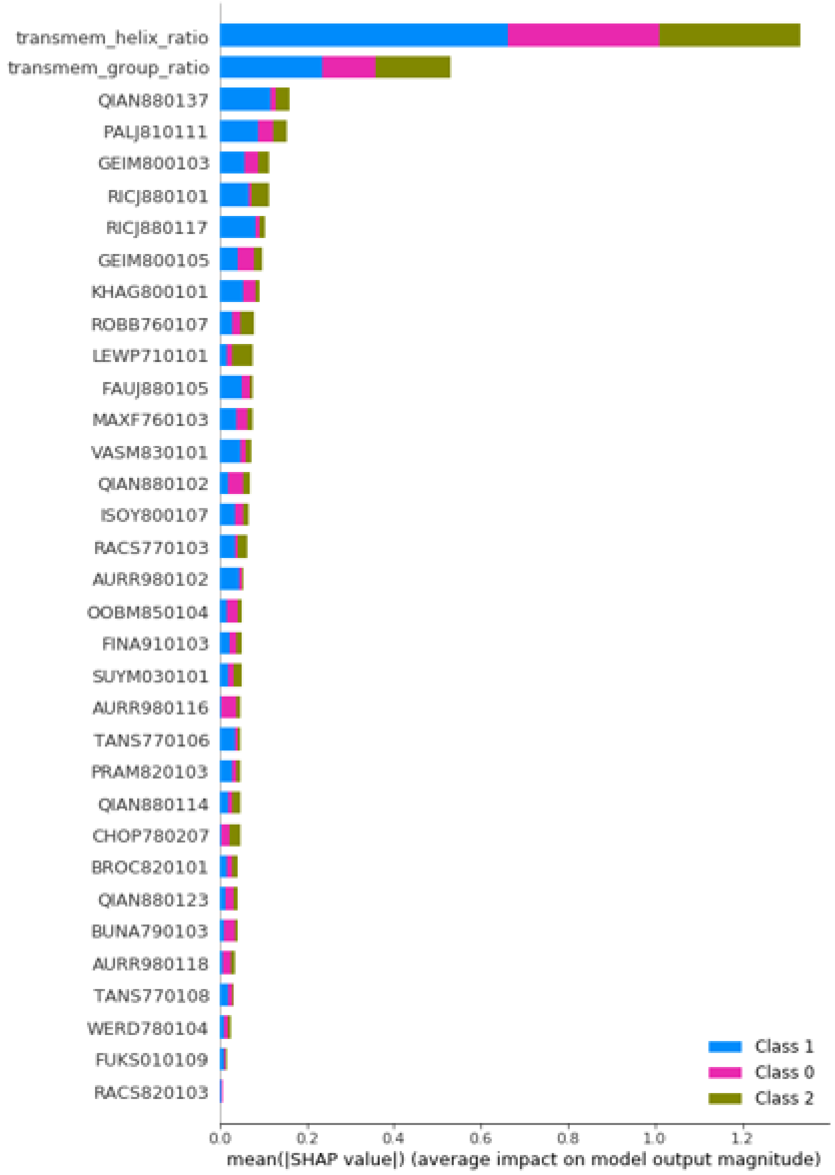

Figure 2

Feature Importance.

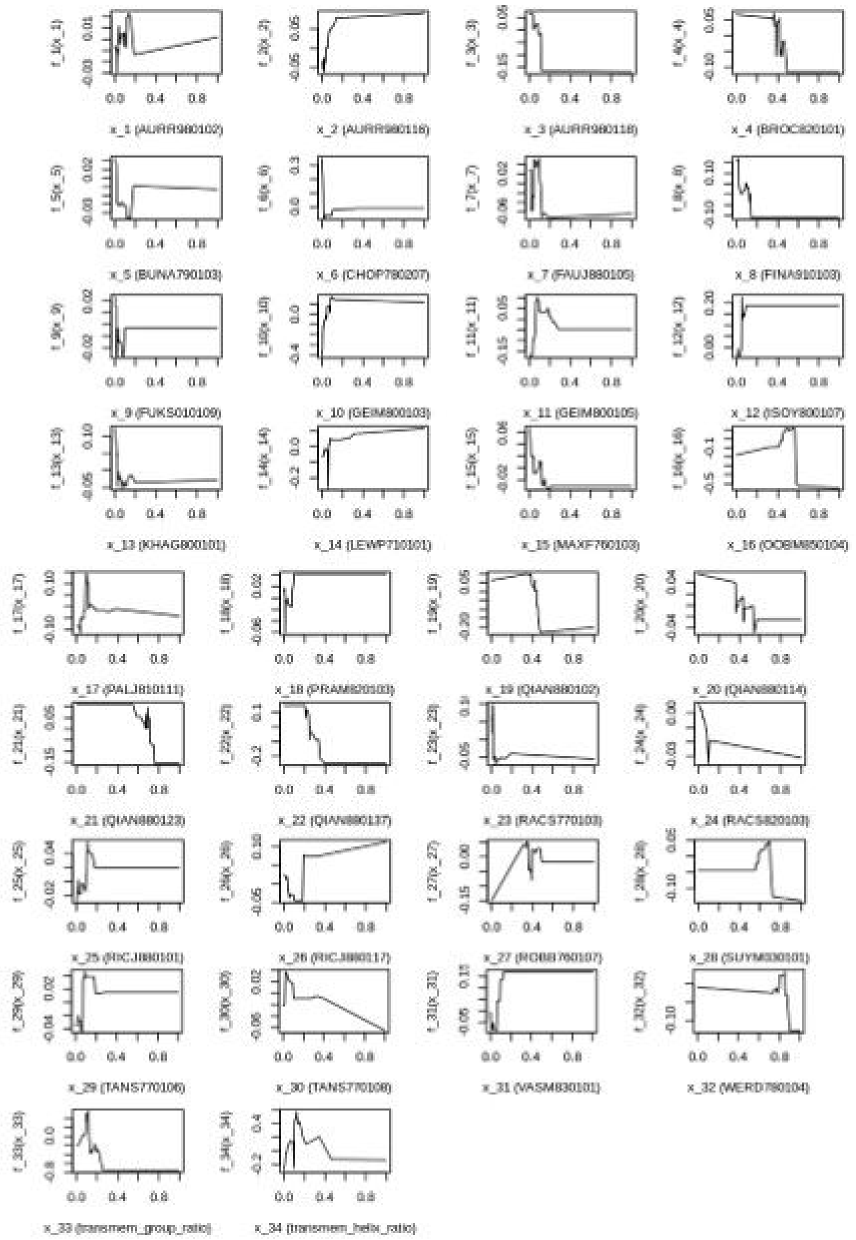

Figure 3

ALE Plots for each feature

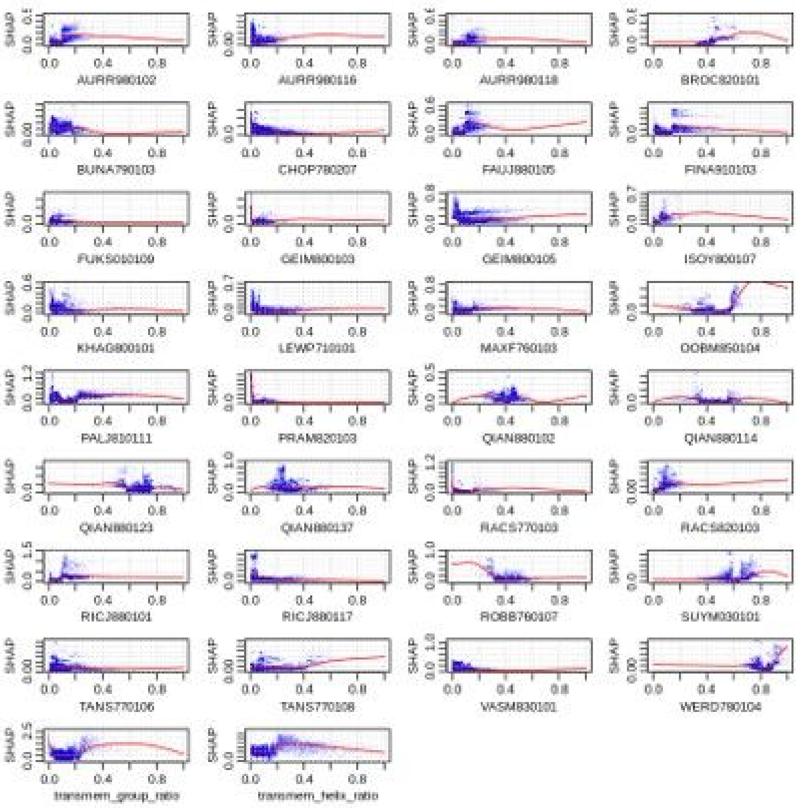

Figure 4

Shapley summary plot. x-axis: original variable value. y-axis: shap value.each blue dot is a row.

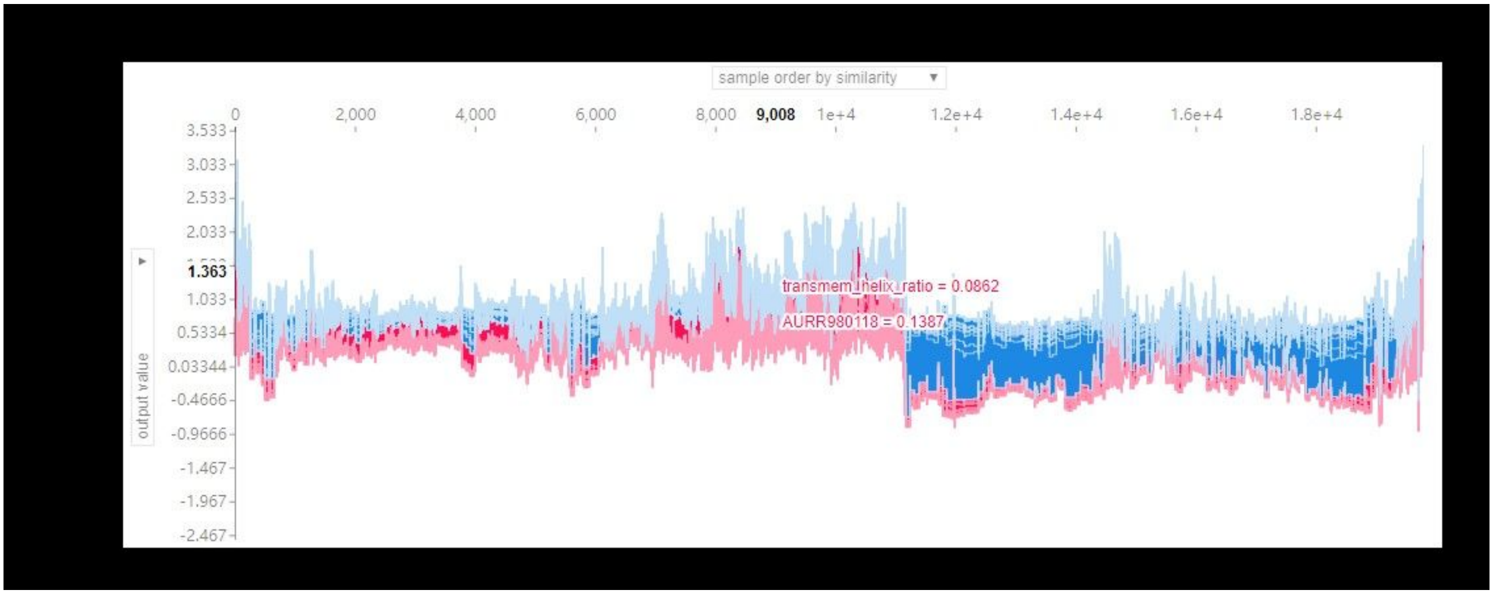

Figure 5

Shapley model explainer

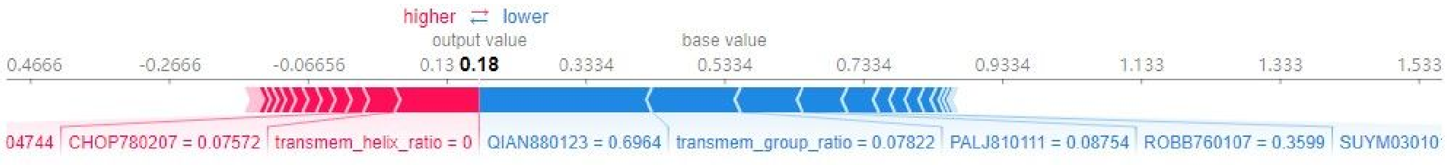

Figure 6

Shapley individual prediction explainer