主論文の要旨

FexSplice: A LightGBM-Based Model for Predicting the Splicing Effect of a Single Nucleotide Variant Affecting the First Nucleotide G of an Exon

（FexSplice: エクソンの最初のヌクレオチドGに影響を与える単一ヌクレオチドバリアントのスプライシング効果を予測するLightGBMモデル）

名古屋大学大学院医学系研究科　総合医学専攻
先端応用医学講座　神経遺伝情報学分野
（指導：大野 欽司　教授）

ATEFEH JOUDAKI

【Background】

The prediction of splicing effects due to single nucleotide variants (SNVs) is an essential aspect of current genomics, especially in understanding the functional implications of pathogenic variants associated with diseases. Splicing—the mechanism through which introns are removed from pre-mRNA to create mature RNA—can be remarkably sensitive to even minor alterations in the nucleotide sequence. In particular, SNVs that occur at the first nucleotide 'G' of an exon (Fex-SNVs) have not been sufficiently studied despite their significant impact on splicing dynamics. Traditional predictive methods often rely on sequence homology or the strength of splicing regulatory elements, overlooking context-specific subtleties. These limitations highlight the need for a more nuanced, machine-learning-based approach. LightGBM, a gradient-boosting framework, has shown promise in various bioinformatics applications but remains largely unexplored in the realm of Fex-SNVs. This study aimed to address this research gap through the introduction of a LightGBM-based model, named 'FexSplice,' designed to predict the splicing effects of Fex-SNVs. FexSplice incorporated features including the polypyrimidine tract length, splicing regulatory elements, and nucleotide composition, aiming to improve the accuracy of splicing effect predictions (Figure 1.)

【Methods】

To develop FexSplice into a model for predicting the splicing effects of Fex-SNVs, a multi-step methodology was implemented. This methodology was broadly categorized into four phases: data collection, feature extraction, model development, and evaluation.

Data Collection: Splicing-affecting Fex-SNVs were first acquired from the Human Gene Mutation Database Pro and ClinVar. Then, an exhaustive literature review was performed to identify Fex-SNVs that affect splicing by experimental evidence. Neutral Fex-SNVs were collected from the dbSNP database, focusing on those with a global minor allele frequency (MAF) between 0.01 and 0.50. This phase involved the curation of a dataset comprising 106 splicing-affecting and 106 neutral Fex-SNVs.

Feature Extraction: We extracted 115 features that encapsulate the underlying biological mechanisms affecting splicing. These features include aspects like the length and nucleotide compositions of the polypyrimidine tract and the strength and position of splicing enhancers and silencers.

Model Development: LightGBM, a gradient-boosting framework, was selected for its performance in various bioinformatics applications. Support Vector Machine (SVM) and Random Forest algorithms were also employed for comparative purposes. Recursive feature elimination techniques were employed to identify the most significant features for accurate predictions.

Model Evaluation: To evaluate the model's performance, we used 10-fold cross-validation. Metrics like accuracy, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUROC) were calculated to evaluate the model's capabilities. The performance of FexSplice was benchmarked against existing models to ascertain its effectiveness.

【Results】

Three machine-learning algorithms—LinearSVC, RandomForest, and LightGBM—were employed to build predictive models. These models were evaluated through several metrics, including Area Under the Receiver Operating Characteristics (AUROC) and Area Under the Precision-Recall Curve (AUPRC) (Figure 2). Additionally, the models were assessed through seven statistical measures—accuracy, precision, recall/sensitivity, specificity, F1 score, Negative Predictive Value (NPV), and Matthews Correlation Coefficient (MCC)—all of which were calculated using a 10-fold cross-validation technique (Table 1).

Compared to LinearSVC and RandomForest, LightGBM achieved the highest AUROC and outperformed the other models in six of the seven statistical measures, with specificity being the exception. The feature importance for each of the 115 features was further inspected using LightGBM and is detailed in Figure 3. To optimize the model, we also conducted a feature elimination exercise using Leave-One-Out Cross-Validation (LOOCV). This led to an interesting observation: while the performance of LinearSVC and RandomForest remained relatively unchanged, the balanced accuracy of LightGBM was maximized when the feature set was narrowed down to just 15. This reduction in features also led to an improvement in all seven statistical measures for LightGBM, with an increase of AUROC from 0.84 ± 0.08 to 0.86 ± 0.08 (Figure 2 and Table 1).

After optimizing LightGBM with a reduced set of 15 features, the model showed improved performance metrics (Table 1), enhancing its capability for predicting splicing effects of Fex-SNVs. This optimized model, which we refer to as FexSplice, had feature importance values that were similar to those of the 115-feature model. These findings indicate that FexSplice, as measured by multiple evaluation metrics, performs consistently in the prediction of splicing effects caused by Fex-SNVs. The results suggest that FexSplice could be useful for future genomics research.

【Discussion】

The development of predictive models for assessing the splicing effects of Fex-SNVs marks a significant advancement in genomics. Among the three models—LinearSVC, RandomForest, and LightGBM—LightGBM emerged as the most promising tool, outperforming the others in almost every evaluation metrics. Specifically, LightGBM

excelled in six out of the seven statistical measures, the only exception being specificity. Although feature overfitting to the dataset could not be eliminated in feature elimination, the elimination of features from 115 to 15 reasonably improved evaluation metrics. The comparison between FexSplice, SpliceAI, and CI-SpliceAI offers an intriguing insight. While SpliceAI and CI-SpliceAI had higher specificity and precision, they lagged considerably in recall. This suggests that although these models are good at reducing false positives, they may miss out on true positives, which could be critical in a clinical setting. Therefore, FexSplice fills an essential gap by achieving a more balanced performance. The FexSplice web service enhances the tool's accessibility. By providing an accessible platform for splicing effect prediction based on genomic coordinates, this service opens up avenues for more extensive genomic research and clinical applications. It is important to acknowledge that, like all other machine-learning models, FexSplice has its limitations. While it excels in a balanced set of statistical measures, there is still room for improvement in specificity. Additionally, the model was developed and validated using a specific dataset, and its performance needs to be validated using an external testing dataset of Fex-SNVs that will be reported in the future.

【Conclusion】

The study presented FexSplice, a machine-learning model, and evaluated its performance using multiple metrics. The model's strength lies in its high scores across multiple evaluation metrics and improved performance upon feature reduction. While comparison with existing models like SpliceAI and CI-SpliceAI highlights FexSplice's balanced capabilities, introducing a user-friendly web service extends its utility and accessibility for broader applications in genomics. Future work should focus on validating FexSplice across diverse datasets and exploring avenues for further optimization.