doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 04. 008

利用 DNA 序列对错义突变的致病性进行预测

镇旭阳[1]，林关宁[1,2]

（1. 上海交通大学生物医学工程学院，上海 200030；2. 上海市精神卫生中心，上海 200030）

【摘要】错义突变的致病性预测在基因组学和临床研究中有重要作用，其中基于计算方法的预测工具已经取得较大进展。现有的工具大多根据功能影响、保守性来对错义突变的致病性进行预测，从 DNA 序列出发进行错义突变致病性预测的工具较少。随着自然语言处理技术在多个生物序列领域的迁移学习和应用，将 DNA 序列作为一种生物语言进行处理并进行基因突变的致病性预测越来越值得探索。该文提出了一个基于预训练的自然语言模型 DNABert 和 DNA 突变序列对错义突变致病性进行预测的深度学习模型 MissenseBert，并且在多个数据集上对 MissenseBert 进行了训练和测试，测试结果说明 MissenseBert 取得了较好的预测效果，证明了利用 DNA 序列预测错义突变的致病性具有可行性。

【关键词】基因突变；错义突变；致病性预测；深度学习

【中图分类号】TP311.13、Q811.4　　【文献标志码】A　　文章编号：1674-1242（2023）04-0381-10

Utilize DNA Sequences to Predict Pathogenicity of Missense Variants

ZHEN Xuyang[1], LIN Guanning[1,2]

(1. The School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China;
2. Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China)

【**Abstract**】Pathogenicity prediction of missense variants is important in genetic research and clinical application. Among current related methods, computational methods have been widely applied and shown outstanding performance. Most computational methods are based on functional impact of alteration or conservation of sequence. Considering some natural language processing (NLP) methods are used in biological sequence tasks by transfer learning, handling DNA sequences as a kind of biological language and predicting the pathogenicity of genetic variants are encouraging. Based on a pretrained NLP model and DNA sequences with altered allele, we propose a deep learning model named MissenseBert to predict the pathogenicity of missense variants. Training and evaluated with multiple datasets, MissenseBert achieves promising performance and illustrate the feasibility of predicting the pathogenicity by DNA sequence.

【**Key words**】Genetic Variant; Missense Variants; Pathogenicity Prediction; Deep Learning

0　引言

基因突变是发生在核苷酸序列上的突变，根据发生的位置和造成的转录翻译后果不同可分为不同的类型，其中错义突变会造成蛋白质中氨基酸的替换[1,2]。这种肽链上的氨基酸替换可能会影响蛋白的正常功能，导致严重的功能障碍和疾病[3-5]。同时，蛋白的表达和功能受到生物细胞中复杂的调控机制的影响，导致错义突变在致病性的表现上存在差异[6,7]。

收稿日期：2022-12-16。

基金项目：国家自然科学基金（No. 81971292, 82150610506），上海市自然科学基金（No. 21ZR1428600）资助项目，上海交通大学医工交叉基因（No. YG2022ZD026）。

作者简介：镇旭阳（1996—），男，硕士研究生，从事机器学习和数据挖掘研究。

通信作者：林关宁，男，教授，博士研究生导师，邮箱（E-mail）：nickgnlin@sjtu.edu.cn。

随着基因测序技术和精准医疗的快速发展，对包括错义突变在内的基因突变进行准确的致病性预测和判断具有重要意义[8,9]。针对基因突变的评估，研究人员开发了非常多的工具。根据实现原理和方法的不同，这些工具可以分为以下 3 类。①根据功能影响进行预测，主要通过发生基因突变后蛋白的理化性质和生物功能的改变来评估致病性[10-12]。②根据 DNA 或蛋白质序列的保守性进行预测，主要通过同源序列比对等方法，评估基因突变后序列的保守性变化程度，从而判断基因突变是否致病[13,14]。③根据集成的思想，利用其他工具的预测结果建立统计学习模型或深度学习模型进行训练，最终输出基因突变的致病性预测分数[15-17]。基于集成方法的工具的输出结果对其他预测工具的预测结果进行综合，其表现往往比基于其他方法的预测工具好[18,19]。对于现有的方法，预测错义突变的致病性需要依赖突变的特征或突变所在位置的同源序列，在缺少突变数据的情况下，预测工具的输出存在较大的限制。同时，大多数工具将全部基因和蛋白上的突变数据作为一个整体数据集进行训练和预测，在部分突变数据匮乏的基因和蛋白上的预测可靠程度存在不确定性。

基因序列由 4 种碱基排列组合得到。蛋白质序列由 21 种氨基酸排列组合得到。生物序列的排列中包含大量的信息，能否通过生物序列排列组合形成的生物语言判断对应的功能令我们感兴趣[20]。通过自然语言处理方法的迁移学习，序列在多个生物学问题的研究中得到了应用[21-25]。受到这些研究的启发，本文提出了一个基于 DNA 语料预训练的自然语言模型——DNABert[26]和发生错义突变后的 DNA 序列，直接对错义突变的致病性进行预测和评估的深度学习模型 MissenseBert，并在多个数据集上对该模型进行了训练和测试，研究从序列出发能否对错义突变的致病性进行准确的预测。

1　模型设计

本文使用调整过的 DNABert 对错义突变进行特征抽取，得到 DNA 序列的深层嵌入表示，并使用 MLP 二分类器对 DNA 序列的深层嵌入表示进行致病性分类，给出致病性概率分数。DNABert 是基于 Bert 在 DNA 语料上预训练得到的模型，其被证明在微调后能够应用到多种与 DNA 相关的下游任务中，但利用 DNABert 进行错义突变序列的致病性预测的效果有待验证[26]。MissenseBert 对收集的错义突变序列进行编码，使用改进后的 DNABert 进行训练和微调。通过 DNABert 编码器模块的多层 Transformer 结构，计算并获取其中不同层的隐状态，再将得到的隐状态向量作为错义突变序列的抽象特征送入全连接层进行二分类。

考虑到每个基因和每个蛋白的表达与功能都存在非常大的差异，MissenseBert 利用每个蛋白上的突变数据建立了一个针对该蛋白的错义突变的预测模型，从而实现对错义突变的准确预测。MissenseBert 的整体框架如图 1 所示。

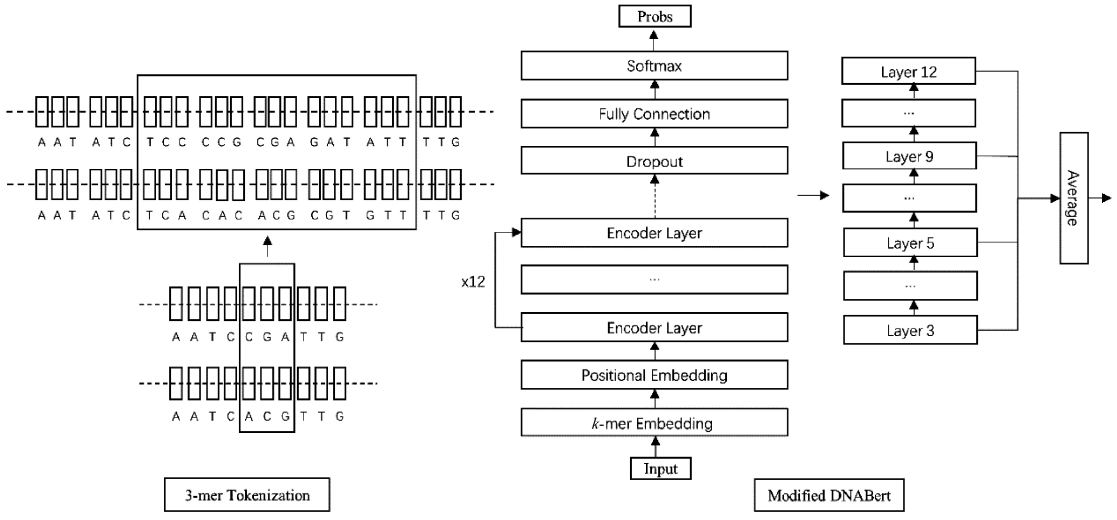


图 1　MissenseBert 的整体框架

Fig. 1　The overall framework of MissenseBert

对于每条 DNA 的突变序列，MissenseBert 先使用 *k*-mer 方法对序列进行分词。例如，对于一个碱基序列{ATTCG}，首先，使用 3-mer 分词后得到{ATT,TTC, TCG}的 3 个词元。使用 3-mer 分词后得到的词本中包括 4^3 个词元，分别表示 4 种碱基的不同组合。其次，MissenseBert 按照词本对词元进行编码，将一条碱基序列映射为向量，从基因序列的“句子”中得到基因序列的“单词”的嵌入表示。最后，MissenseBert 利用 DNABert 对错义突变序列的深层嵌入表示进行学习，DNABert 通过 12 层 Transfomer 编码器[27]对基因序列的编码特征和上下文信息进行从低层次到高层次的抽取。Transformer 中的多头注意力机制能够捕捉到序列中不同位置的上下文信息，并且 DNABert 已经在大量 DNA 语料的基础上进行了预训练，因此 DNABert 能够捕捉到错义突变序列不同位置的碱基之间的相关性。其中，对于一条突变序列 S 和第 i 个注意力有

$$\begin{aligned}&\text{MultiHeadAttention}(S)\\&=\text{Concat}(\text{attention}_1,\text{attention}_2,\cdots,\text{attention}_h)W^O\end{aligned}$$

$$\text{attention}_i=\text{softmax}\left(\frac{SW_i^QSW_i^{K^T}}{\sqrt{d_k}}\right)\cdot SW_i^V$$

式中，W^O、W_i^Q、W_i^K 和 W_i^V 是训练过程中学习的参数，d_k 是序列查询向量的维度，S 的隐状态通过计算序列中每个词之间的注意力分数加权到 W_i^V 得到。

Bert 和 DNABert 的最初实现不同，MissenseBert 选择将 12 层 Transformer 中的 4 层（第 3、5、9、12 层）的输出向量进行平均取值，作为输出序列的最终抽象表示特征，最后将结合不同层次的序列特征送入包括 Dropout 层、全连接层和 softmax 层在内的 MLP 分类器进行分类，得到每个错义突变的致病性概率分数，并根据设定的阈值得到一个错义突变是否致病的结论。

2　模型实现

2.1　数据集

我们使用 EVE[28]提供的错义突变数据集作为主要使用的数据集，并在 Clinvar 数据库[29]和人类标准基因组中验证 MissenseBert 的效果。EVE 是一个结合序列保守性和深度学习方法的错义突变致病性评估工具，其取得了当前最佳的预测表现，并且具有和实验方法几乎一致的预测表现。EVE 提供了超过 3 000 个蛋白上的错义突变，并且对每个错义突变都给出了是否致病的标签。鉴于 EVE 最佳的预测表现，我们利用 EVE 的错义突变数据对 MissenseBert 进行训练。移除 EVE 数据集中和人类标准基因组序列 GRCh38 不一致的突变位点，并以此为依据将致病的突变标签设为 1，将不致病的突变标签设为 0。最终，我们使用了 2 207 个蛋白上的 62 249 784 个错义数据用于训练和测试，其中蛋白 HERC2_HUMAN 上有最多的 272 710 个错义突变，蛋白 TWST1_HUMAN 上有最少的 167 个错义突变。我们收集了 Clinvar 数据库中的错义突变，由于 Clinvar 数据库中的错义突变以碱基位点的形式给出，而 EVE 中的错义突变以蛋白位点的形式给出，因此我们在两者之间进行了转换。另外，根据 Clinvar 中的评估标准，我们移除了其中可信度较低的错义突变以确保验证的准确性，同时将临床显著性为“良性”和“可能良性”的突变标签设为 0，将“致病”和“可能致病”的突变标签设为 1。此外，我们从 GRCh38 版本的人类标准基因组中提取了和错义突变对应的野生型参考序列。Clinvar 和标准基因组的数据被作为独立的测试数据集，我们利用这两个独立的测试数据集对 MissenseBert 的预测效果进行了进一步验证。

2.2　评价指标

我们使用常见的二分类指标来评估 MissenseBert 的表现，包括准确率（ACC）、精确率（Precision）、召回率（Recall）、F1 得分（f_1）和 Matthews 相关系数（MCC）。其中，F1 得分和 Matthews 相关系数能够反映在正负样本数量不均衡的情况下，分类器在正负两类样本上的综合表现。

$$\text{ACC}=\frac{\text{TP}+\text{TN}}{\text{T}+\text{F}}$$

$$\text{Precison}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$f_1=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{MCC}=\frac{\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})\cdot(\text{TP}+\text{FN})\cdot(\text{TN}+\text{FP})\cdot(\text{TN}+\text{FN})}}$$

式中，TP、TN、FP、FN 分别表示在分类器预测结果中真阳性、真阴性、假阳性和假阴性的样本数量。

2.3　数据处理

对于 EVE 中的错义突变数据，每个蛋白上的错义

突变数据都按照 8∶2 的比例划分为训练验证数据集和测试数据集，对训练验证数据集也按照 8∶2 的比例划分为训练数据集和训练过程中的验证数据集。训练验证数据集用于模型训练过程中的调参，而测试数据集用于评估预测的表现。其中，训练数据集、训练过程中的验证数据集和测试数据集之间没有重复的错义突变。

对于一个错义突变，我们获取了包括错义突变在内的 101 个碱基长度的序列作为模型输入，肽链上的 1 个氨基酸的替换对应核苷酸链上最长 3 个碱基的改变，我们在 3 个碱基的左右各拼接 48 个标准基因组中的参考序列碱基作为突变后的序列，随后将 101 个碱基长度的序列按照 3-mer 的方式进行分词。

2.4　训练参数

我们使用 AdamW 训练器[30]对 MissenseBert 进行训练。由于 MissensesBert 只需要在预训练模型 DNABert 的基础上进行微调，因此我们设置了较小的学习率 10e-5，将 dropout 率设置为 0.1，微调的迭代次数设置为 5 次。

我们在一块 NVIDIA GeForce RTX 3090 上使用 Pytorch[31]进行训练，使用的 Pytorch 版本是 1.10.1，CUDA 版本是 11.4。

3　模型评估

3.1　整体分类性能

利用每个蛋白上的数据训练模型后，我们一共得到了 2 207 个在不同蛋白错义突变上分别训练的预测模型。我们使用每个蛋白对应的测试数据集对错义突变进行预测，统计了 6 项分类评价指标。对于全部的 2 207 个蛋白，MissenseBert 的预测表现较好，其中平均 AUC 值为 0.800，MCC 值为 0.546，这说明在部分蛋白的预测上较为准确，同时在部分蛋白的预测上出现了预测不准确和不平衡的情况。我们按照 AUC 指标对 MissenseBert 在各蛋白上的预测表现进行了降序排列，分别统计了前 1 500 个、前 1 000 个、前 500 个蛋白的平均预测表现，如表 1 所示。

其中，在前 500 个蛋白上，MissenseBert 的表现非常好，AUC 值达到 0.889，MCC 值达到 0.773，说明在前 500 个表现最好的蛋白上，MissenseBert 对致病的和不致病的错义突变都能做出准确的预测。而在前 1 000 个表现最好的蛋白预测结果中，MissenseBert 的 AUC 值也达到了 0.856，MCC 达到 0.709，说明在接近一半的蛋白上，MissenseBert 能对错义突变的致病性做出较为准确的预测，如图 2 所示。

表 1　MissenseBert 的整体预测表现

Tab. 1　The overall performance of MissenseBert

	ACC	AUC	f_1	MCC	Precision	Recall
所有蛋白	0.800	0.768	0.787	0.546	0.793	0.800
前 1 500 个蛋白	0.836	0.830	0.835	0.661	0.839	0.836
前 1 000 个蛋白	0.859	0.856	0.859	0.709	0.863	0.859
前 500 个蛋白	0.891	0.889	0.891	0.773	0.895	0.891

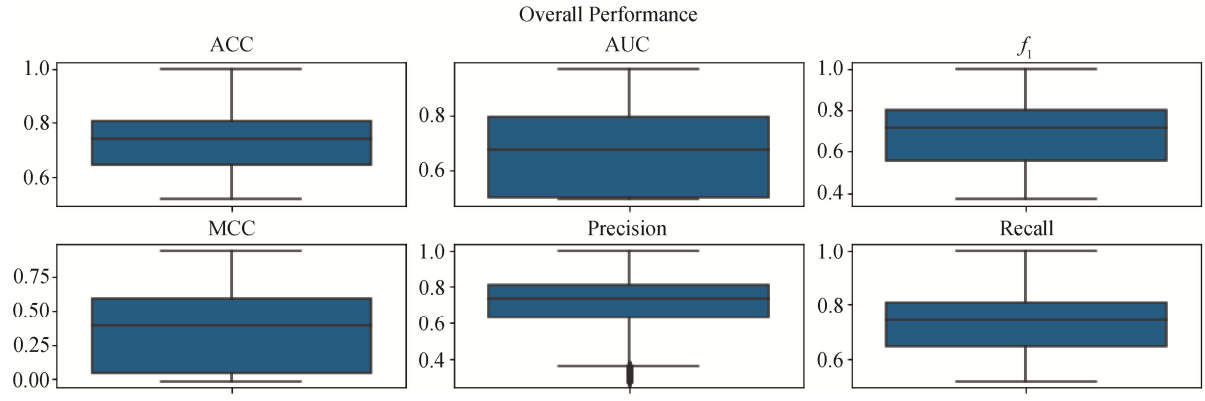


图 2　MissenseBert 在预测效果最好的前 1 000 个蛋白上的表现

Fig. 2　The prediction performance of top 1 000 proteins by MissenseBert

MissenseBert 预测表现排名前 5 位的蛋白分别是 RFOX1_HUMAN、PRIC2_HUMAN、ELF2_HUMAN、PAX2_HUMAN、TPO_HUMAN，它们的 AUC 值都在 0.95 以上，如表 2 所示，说明在这 5 个蛋白上 MissenseBert 的预测结果和突变标签基本一致。

为了更好地确认 MissenseBert 的预测效果，我们选择 AUC 值排名第 100 位的蛋白 NSD1_HUMAN 和排名第 500 位的蛋白 WDR37_HUMAN，对这两个蛋白的分类结果进行检查。我们得到了预测结果的混淆矩阵和 ROC 曲线，如图 3 所示。可以看到，无论是对于这两个蛋白上的致病突变还是良性突变，MissenseBert 都能进行较为准确的预测。

表 2　MissenseBert 预测表现排名前 5 位的蛋白

Tab. 2　The 5 proteins with best prediction performance

蛋白	ACC	AUC	f_1	MCC	Precision	Recall
RFOX1_HUMAN	0.983	0.985	0.983	0.954	0.983	0.983
PRIC2_HUMAN	0.970	0.975	0.970	0.917	0.972	0.970
ELF2_HUMAN	0.981	0.972	0.981	0.943	0.981	0.981
PAX2_HUMAN	0.960	0.965	0.961	0.913	0.962	0.960
TPO_HUMAN	0.950	0.964	0.951	0.884	0.957	0.950

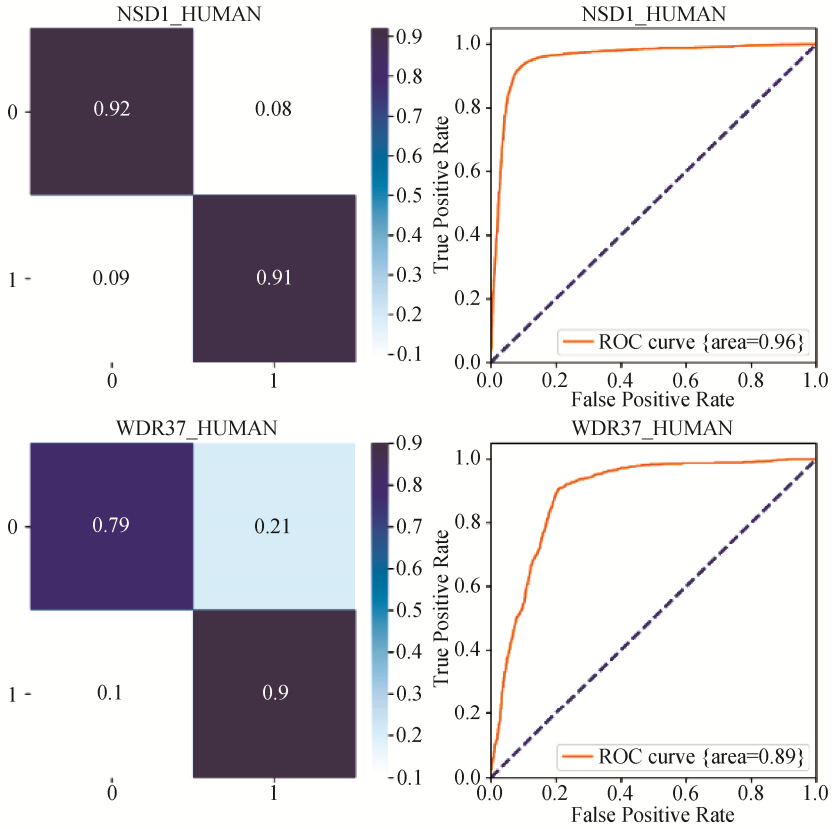


图 3　NSD1_HUMAN 和 WDR37_HUMAN 的混淆矩阵和 ROC 曲线

Fig. 3　The confusion matrices and ROC curves of NSD1_HUMAN and WDR37_HUMAN

3.2　利用野生型参考序列进行验证

当把标准基因组上的野生型参考序列作为输入时，理论上突变预测工具应该将输入序列预测为不致病，否则工具对正常序列的预测不可靠。在 EVE 数据集中，我们选择了参考序列位点最多的 30 个蛋白，将每个参考位点和对应的野生型参考序列作为输入，同时将所有输入序列的标签设置为不致病，验证 MissenseBert 的预测表现，如图 4 所示。

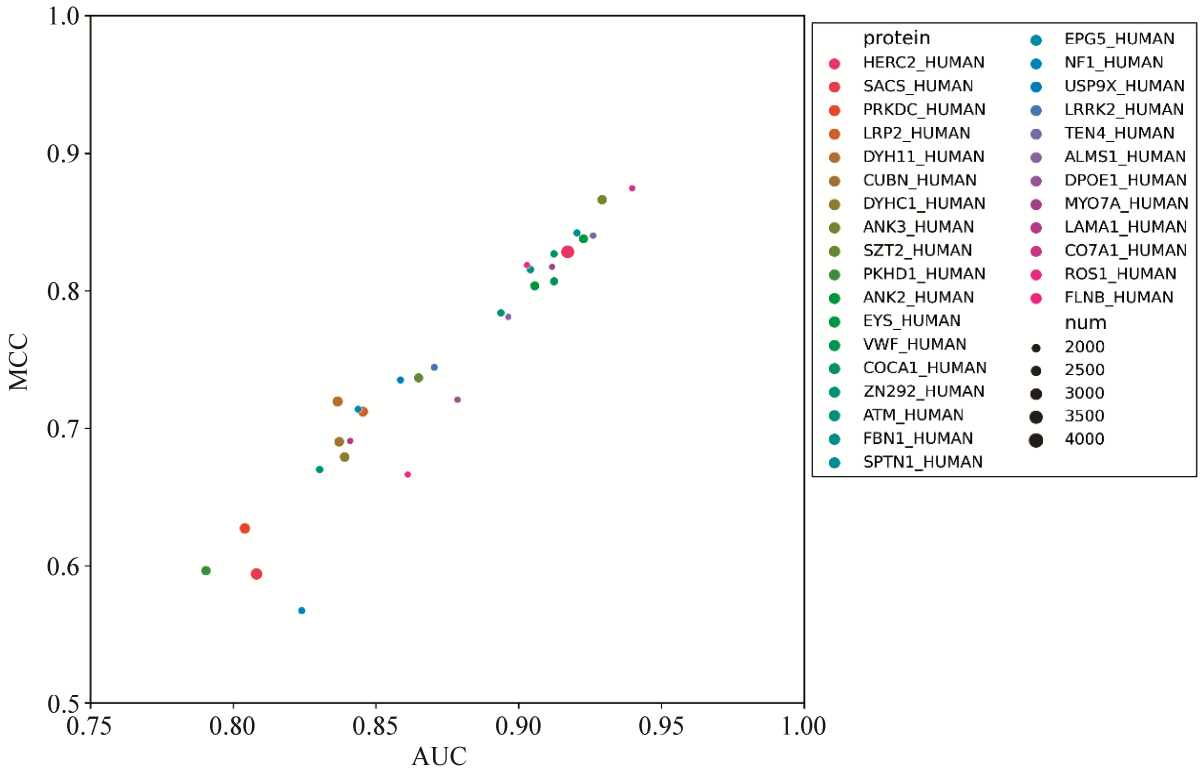


图 4　MissenseBert 在野生型参考序列上的预测表现

Fig. 4　The prediction performance of MissenseBert with wild type genome sequences

尽管在训练数据集中不包括参考序列，但是 MissenseBert 预测的评价指标仍然较好，说明大部分野生型参考序列被预测为不致病，MissenseBert 能够很好地区分出参考序列和突变序列。

3.3　在 Clinvar 数据库中验证

Clinvar 数据库中收集了各种基因突变，并且标记了每条基因突变的致病性。尽管有文献指出 Clinvar 中的一些基因突变的标签可能存在偏差[32]，但 Clinvar 仍然是基因突变预测领域最重要的数据库之一，被广泛应用于基于监督学习的预测工具的开发。我们将 Clinvar 中出现的错义突变作为独立的测试数据集，来检验 MissenseBert 在 EVE 数据集之外的实际突变数据上的预测表现。

我们选择了 30 个蛋白上的错义突变来进行预测。和其他蛋白相比，这些蛋白在 Clinvar 数据库中有最多的错义突变收录，而其他蛋白在 Clinvar 中收录的数量较少，所以我们认为在这些蛋白上进行验证更具有说服力。

可以看到，对于独立的 Clinvar 数据集，MissenBert 可以对其中的错义突变做出较为准确的致病性预测。对于 30 个错义突变条数最多的蛋白，MissenseBert 预测结果的 AUC 值均在 0.75 以上，MCC 值均在 0.5 以上（见图 5）。

3.4　和其他预测工具的对比

我们选择了 5 个应用广泛的错义突变致病性预测工具即 MutationTaster[12]、FATHMM[11]、M-CAP[16]、REVEL[17]和 CADD[15]作为对比工具，并将 3.3 节中的 30 个蛋白的 Clinvar 数据和 MissenseBert 进行预测表现的对比。我们使用 AUC 的平均值来评估各个工具的综合预测表现，其中 MissenseBert 的 AUC 平均值为 0.854，和 CADD 的表现基本持平，优于 MutationTaster 和 FATHMM，但是劣于 M-CAP 和 REVEL（见图 6）。这些对比工具在训练过程中可能使用了 Clinvar 数据，同时依赖较多的特征，而 MissenseBert 在只使用序列作为输入的情况下仍取得了和对比工具非常接近的表现。

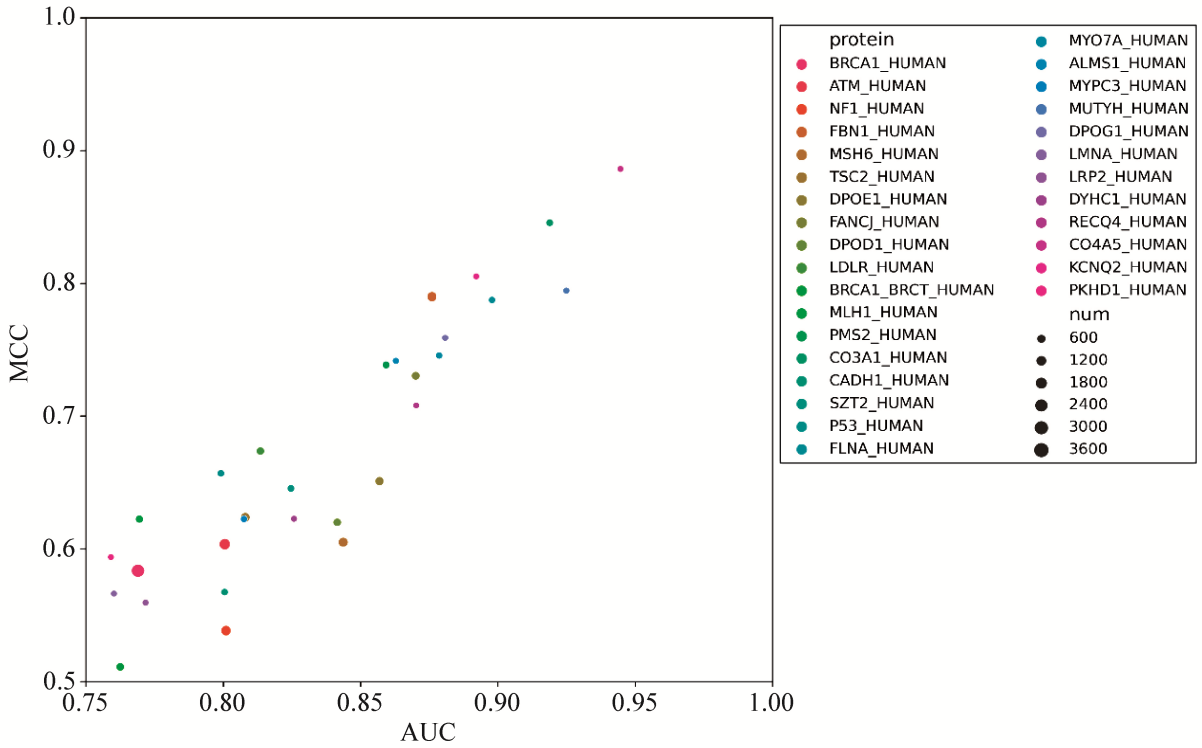


图 5　MissenseBert 在 Clinvar 数据集上的预测表现

Fig. 5　The prediction performance of MissenseBert with Clinvar dataset

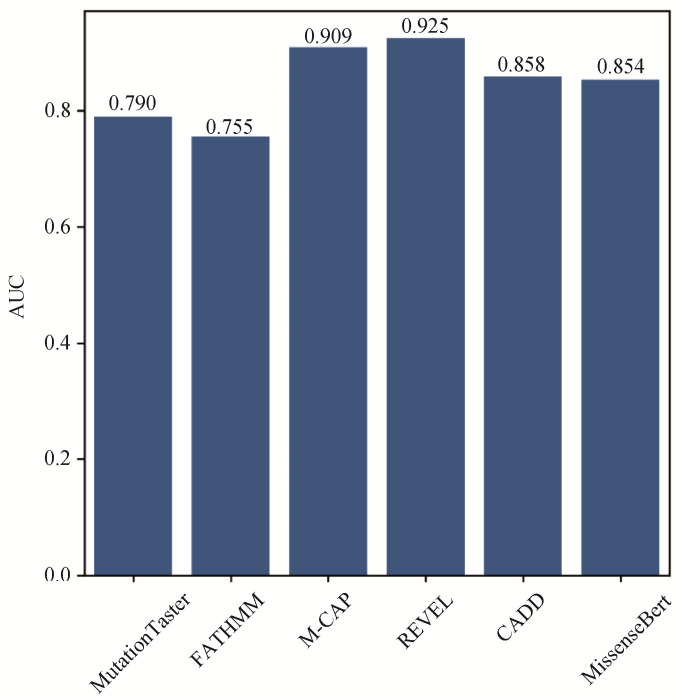


图 6　MissenseBert 和其他工具的对比

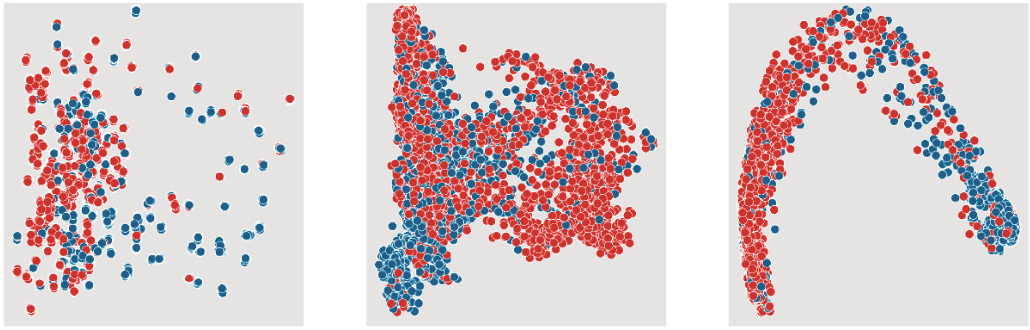
Fig. 6　Comparison of performance between MissenseBert and other tools

3.5　序列编码向量可视化

为了直观地说明 MissenseBert 中特征抽取和微调的有效性，我们选择将 P53_HUMAN 蛋白上的错义突变编码向量进行降维和可视化。我们得到 P53_HUMAN 蛋白上的错义突变在使用 DNABert 进行特征抽取前后及微调前后的分布，并使用 PCA 方法[33]进行降维后观察两者的区别。

在不使用 DNABert 进行特征抽取的情况下，致病性和良性突变混杂在一起［见图 7（a）］。

同样地，在不微调的情况下，突变的分布也不易区分［见图 7（b）］。而在使用 DNABert 进行编码和微调后，两种突变的分布出现了较为明显的聚簇［见图 7（c）］。这说明 MissenseBert 中的 12 层 Transformer 编码器特征抽取和微调过程的有效性——致病突变和非致病突变的编码向量被较好地区分开来。和原始的编码向量相比，特征抽取和微调后的编码向量更适合使用 MLP 分类器进行分类，这说明 MissenseBert 中基于 DNABert 的特征抽取和微调取得了较好的效果，也解释了 MissenseBert 取得较好的预测表现的原因。



（a）不使用 DNABert 进行特征抽取　（b）不微调　（c）使用 DNABert 进行编码和微调

图 7　P53_HUMAN 蛋白上错义突变编码对比

Fig. 7　The comparison of variant encoding in P53_HUMAN

4　结语

随着下一代基因测序技术的持续发展，个性化医疗和精准医疗的应用不断出现。作为生命活动的起点，基因组学的研究非常关键，其中预测基因突变和错义突变的致病性对精准医疗方案的开发和应用具有重要意义。现有的预测方法往往依赖多种特征和工具，尽管基于特征和集成的工具能够取得较好的预测效果，但是在特征缺失的情况下，它们不能对错义突变的致病性进行有效而可靠的预测。随着深度学习技术特别是自然语言处理技术在生物序列上的迁移应用，根据碱基和氨基酸的排列来研究序列有关的生物问题有了越来越多令人振奋的探索。

MissenseBert 受到这些研究及 Bert 和 DNABert 的启发，结合预训练和自注意力的设计思想，将 DNA 突变序列进行合理编码和特征抽取，通过深度学习分类器预测错义突变的致病性。经过在多个数据集上的验证，MissenseBert 在错义突变致病性的预测上取得了和当前最优方法非常接近的预测表现，在 Clinvar 数据库和标准基因组的测试数据集中也能够较准确地预测。MissenseBert 的表现说明在不依赖其他特征的情况下，仅使用 DNA 突变的序列来预测错义突变的致病性也具有可行性。

MissenseBert 在部分基因和蛋白上取得了非常好的预测效果，而在小部分基因和蛋白上的预测表现却并不理想，这可能是因为不同基因上的错义突变受到的内部调控和功能机制存在较大差异。这种基因之间的差异表明对不同基因上的错义突变可能需要采取不同的预测方法，这为设计一个通用的突变预测工具带来了困难。为了取得更精确的错义突变致病性预测结果，对每个基因建立单独的模型成为该研究领域的新思路，同时日益丰富的计算资源也为这种新思路提供了有力的支撑。然而，单独的监督学习模型需要每个基因上都有足够的错义突变数据进行训练，这和一些基因上训练数据匮乏之间存在矛盾。MissenseBert 通过使用 EVE 模型预测得到的错义突变数据进行训练和验证，而如何克服真实的错义突变标签不足的问题并利用基于序列的深度学习方法对基因突变进行致病性预测仍然值得探索和研究。

参考文献

[1] LEK M, KARCZEWSKI K J, MINIKEL E V, *et al*. Analysis of protein-coding genetic variation in 6706 humans[J]. **Nature**, 2016, 536(7616): 285-291.

[2] BOYCOTT K M, VANSTONE M R, BULMAN D E, *et al*. Rare-disease genetics in the era of next-generation sequencing: discovery to translation[J]. **Nature Reviews Genetics**, 2013, 14(10): 681-691.

[3] TENNESSEN J A, BIGHAM A W, O'CONNOR T D, *et al*. Evolution and functional impact of rare coding variation from deep sequencing of human exomes[J]. **Science**, 2012, 337(6090): 64-69.

[4] MACARTHUR D G, MANOLIO T A, DIMMOCK D P, *et al*. Guidelines for investigating causality of sequence variants in human disease[J]. **Nature**, 2014, 508(7497): 469-476.

[5] STEFL S, NISHI H, PETUKH M, *et al*. Molecular mechanisms of disease-causing missense mutations[J]. **Journal of Molecular Biology**, 2013, 425(21): 3919-3936.

[6] THUSBERG J, VIHINEN M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods[J]. **Human Mutation**, 2009, 30(5): 703-714.

[7] MIOSGE L A, FIELD M A, SONTANI Y, *et al*. Comparison of predicted and actual consequences of missense mutations[J]. **Proceedings of the National Academy of Sciences**, 2015, 112(37): E5189-E5198.

[8] SQUASSINA A, MANCHIA M, MANOLOPOULOS V G, *et al*. Realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice[J]. **Pharmacogenomics**, 2010, 11(8): 1149-1167.

[9] GINSBURG G S, WILLARD H F. Genomic and personalized medicine: foundations and applications[J]. **Translational Research**, 2009, 154(6): 277-287.

[10] KUMAR P, HENIKOFF S, NG P C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm[J]. **Nature Protocols**, 2009, 4(7): 1073-1081.

[11] SHIHAB H A, GOUGH J, COOPER D N, *et al*. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models[J]. **Human Mutation**, 2013, 34(1): 57-65.

[12] SCHWARZ J M, RöDELSPERGER C, SCHUELKE M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations[J]. **Nature Methods**, 2010, 7(8): 575-576.

[13] DAVYDOV E V, GOODE D L, SIROTA M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++[J]. **PLoS Comput Biol**, 2010, 6(12): e1001025.

[14] SIEPEL A, BEJERANO G , PEDERSEN J S, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes[J]. **Genome Res**, 2005, 15(8): 1034-1050.

[15] KIRCHER M, WITTEN D M, JAIN P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants[J]. **Nature Genetics**, 2014, 46(3): 310-315.

[16] JAGADEESH K A, WENGER A M, BERGER M J, *et al*. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity[J]. **Nature Genetics**, 2016, 48(12): 1581-1586.

[17] IOANNIDIS N M, ROTHSTEIN J H, PEJAVER V, *et al*. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants[J]. **The American Journal of Human Genetics**, 2016, 99(4): 877-885.

[18] LI J, ZHAO T, ZHANG Y, *et al*. Performance evaluation of pathogenicity-computation methods for missense variants[J]. **Nucleic Acids Research**, 2018, 46(15): 7793-7804.

[19] GRIMM D G , AZENCOTT C-A, AICHELER F, *et al*. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity[J]. **Human Mutation**, 2015, 36(5): 513-523.

[20] SEARLS D B. The language of genes[J]. **Nature**, 2002, 420(6912): 211-217.

[21] CHO K, VAN MERRIENBOER B, GULCEHRE C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1724–1734.

[22] SHEN Z, BAO W, HUANG D S. Recurrent neural network for predicting transcription factor binding sites[J]. **Scientific Reports**, 2018, 8(1): 15270.

[23] AVSEC Ž, AGARWAL V, VISENTIN D, *et al*. Effective gene expression prediction from sequence by integrating long-range interactions[J]. **Nature Methods**, 2021, 18(10): 1196-1203.

[24] BASITH S, MANAVALAN B, SHIN T H, *et al*. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree[J]. **Computational and Structural Biotechnology Journal**, 2018, 16: 412-420.

[25] BRANDES N, OFER D, PELEG Y, *et al*. ProteinBERT: a universal deep-learning model of protein sequence and function[J]. **Bioinformatics**, 2022, 38(8): 2102-2110.

[26] JI Y, ZHOU Z, LIU H, *et al*. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome[J]. **Bioinformatics**, 2021, 37(15): 2112-2120.

[27] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//31st Conferenceon Neural Information Processing Systems (NIPS 2017). Long Beach, USA: Curran Associates Inc,2017: 6000–6010.

[28] CAO Y, GEDDES T A, YANG J Y H, *et al*. Ensemble deep learning in bioinformatics[J]. **Nature Machine Intelligence**, 2020, 2(9): 500-508.

[29] LANDRUM M J, LEE J M, BENSON M, *et al*. ClinVar: public archive of interpretations of clinically relevant variants[J]. **Nucleic Acids Research**, 2016, 44(D1): D862-D868.

[30] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]// International Conference on Learning Representations (ICLR 2019). New Orleans, USA: OpenReview.net, 2019:1-18.

[31] PASZKE A, GROSS S, MASSA F, *et al*. Pytorch: an imperative style, high-performance deep learning library[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2019: 8026–8037.

[32] SHAH N, HOU Y-CC, YU H C, *et al*. Identification of misclassified clinvar variants via disease population prevalence[J]. **The American Journal of Human Genetics**, 2018, 102(4): 609-619.

[33] MAĆKIEWICZ A，RATAJCZAK W. Principal components analysis (PCA)[J]. **Computers & Geosciences**, 1993, 19(3): 303-342.