

基于SVM集成学习分析TCR与肽的特异性结合

白宁仪*, 李国洁

河南科技大学数学与统计学院, 河南 洛阳

收稿日期: 2024年7月1日; 录用日期: 2024年7月26日; 发布日期: 2024年8月1日

摘要

本文提出了一种两阶段的方法来预测肽与T细胞受体(TCR)的特异性结合,旨在通过逐步优化预测过程来提高准确性。在第一阶段,我们采用堆叠式自动编码器对肽和TCR序列进行数值嵌入,特别是关注TCR β 链的CDR3区域,这是肽识别的关键决定因素。通过Atchley因子编码,我们将氨基酸的生化特性转换为数字矩阵,并利用无监督学习捕捉序列的关键特征。实验结果表明,自动编码器能够高度忠实地重建原始序列,验证了数值嵌入的有效性。在第二阶段,我们基于第一阶段生成的数值编码,构建了一个集成学习模型来预测肽与TCR的特异性结合。该模型结合了不同内核的支持向量机(SVM)作为基学习器,并通过堆叠法集成它们的预测结果,以提高模型的泛化能力和捕捉基学习器之间的互补性。实验结果显示,集成学习模型的性能显著优于单一的SVM模型,其ROC值的提升,表明集成学习在预测肽与TCR特异性结合方面具有更高的准确性。本文的创新点在于结合了自动编码器的数值嵌入技术和集成学习的预测模型,不仅提高了预测的准确性,还为生物信息学领域的序列分析提供了新的方法论。

关键词

集成学习, 特异性结合, SVM, 自动编码器

Analysis of Specific Binding between TCR and Peptides Based on SVM Ensemble Learning

Ningyi Bai*, Guojie Li

School of Mathematics and Statistics, Henan University of Science and Technology,
Luoyang Henan

Received: Jul. 1st, 2024; accepted: Jul. 26th, 2024; published: Aug. 1st, 2024

*通讯作者。

文章引用: 白宁仪, 李国洁. 基于SVM集成学习分析TCR与肽的特异性结合[J]. 应用数学进展, 2024, 13(8): 3618-3624.
DOI: 10.12677/aam.2024.138344

Abstract

This article proposes a two-stage approach to predict the specific binding of peptides to T cell receptors (TCRs), aiming to improve accuracy by gradually optimizing the prediction process. In the first stage, we use a stacked autoencoder to numerically embed peptides and TCR sequences, particularly focusing on the CDR3 region of the TCR β chain, which is a key determinant of peptide recognition. By encoding the Atchley factor, we transform the biochemical characteristics of amino acids into a numerical matrix and use unsupervised learning to capture key features of the sequence. The experimental results show that the autoencoder can highly faithfully reconstruct the original sequence, verifying the effectiveness of numerical embedding. In the second stage, we constructed an ensemble learning model based on the numerical encoding generated in the first stage to predict the specific binding of peptides to TCR. This model combines support vector machines (SVM) with different kernels as base learners and integrates their prediction results through stacking to improve the model's generalization ability and capture the complementarity between base learners. The experimental results show that the performance of the ensemble learning model is significantly better than that of a single SVM model, and the improvement in its ROC value indicates that ensemble learning has higher accuracy in predicting peptide TCR specific binding. The innovation of this article lies in the combination of numerical embedding technology of autoencoders and prediction models of ensemble learning, which not only improves the accuracy of prediction, but also provides a new methodology for sequence analysis in the field of bioinformatics.

Keywords

Ensemble Learning, Specific Binding, SVM, Autoencoder

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在目前的免疫治疗领域, 许多种策略已经取得了显著的临床成果, 改变了许多恶性肿瘤的治疗范式, 包括免疫检查点抑制剂和 CAR-T 细胞疗法。然而, 尽管这些治疗方法取得了巨大的进步, 但仍然存在许多患者对治疗的抵抗或耐受性, 部分原因在于肿瘤细胞的多样性和免疫逃逸机制的存在。所以免疫治疗领域对于 TCR 与抗原的特异性结合极为重视, 因为这决定了治疗药物或疫苗是否能够准确地识别并攻击目标病原体或异常细胞, 从而提高治疗效果和安全性。通过确保治疗药物或疫苗只与目标抗原特异结合, 可以避免自身免疫反应的发生, 并为个体化治疗策略的设计提供重要依据。

TCR 是免疫系统中的一个重要膜上受体, 它的作用是识别由 MHC 呈递的抗原肽, 并激活 T 细胞, 从而启动免疫反应。每个 TCR 都是由两条不同的多肽链(α 链和 β 链)组成, 其多样性是由基因重排和随机突变产生的。这种多样性使得 TCR 能够识别各种各样的抗原, 并在识别病原体或异常细胞方面具有高度的特异性。TCR 与抗原的特异性结合过程是免疫系统启动免疫应答的关键一步。MHC 分子在抗原呈递中发挥重要作用, 将抗原肽展示给 TCR [1]。通过分析 TCR 与抗原的结合机制, 我们可以更好地理解免疫系统的复杂性。这为开发预测模型提供了动力, 利用这些模型可以帮助研究人员更快、更准确地预测 TCR 与抗原的结合特性。但是传统的实验方法在研究 TCR 与抗原的结合特异性时存在一些局限性。

那些方法通常需要大量的时间和资源, 包括提取和纯化大量的 T 细胞和抗原, 以及进行体外实验来评估它们之间的相互作用。这不仅耗时耗力, 而且成本高昂, 限制了对大规模 TCR 与抗原结合的研究。此外, 由于 TCR 与抗原的结合是高度特异的, 因此需要对许多不同的 TCR 和抗原进行实验才能获得全面的了解。这增加了实验的复杂性和成本, 并且可能无法覆盖所有可能的 TCR-抗原对。因此, 传统实验方法可能无法提供对 TCR 与抗原结合特异性的全面评估。所以我们需要采用一种高效且准确的预测模型可以弥补这些实验方法的局限性。这种预测模型可以快速而准确地评估大量的 TCR-抗原对, 从而为研究人员提供了更全面的理解, 并加速新疗法和疫苗的开发过程。

支持向量机(Support Vector Machine, SVM)是一种监督学习算法, 其主要目标是构建一个能够将数据点划分到不同类别的决策边界[2], 并且使得该边界与各类别数据点的间隔尽可能大。具体来说, SVM 在生物信息学中的应用包括但不限于: 蛋白质结合预测、基因功能预测。通过基于 SVM 算法的预测模型, 我们可以利用大量的已知结合数据来训练模型, 从而提高预测的准确性和可靠性。这种方法将为免疫学研究提供一种全新的工具和方法, 有望推动该领域的进步。通过准确预测 TCR 与抗原的结合特异性, 我们可以更好地理解免疫应答的机制, 揭示免疫系统如何识别和攻击外来病原体以及异常细胞。这项研究还可以帮助筛选出具有高结合亲和力的 TCR-抗原对, 从而加速药物研发过程。此外, 通过预测免疫原和设计个性化的免疫治疗方案, 我们可以更有效地治疗癌症、自身免疫性疾病等疾病, 为患者带来良好的治疗效果。因此, 基于 SVM 算法开发的 TCR 与抗原结合特异性预测模型具有重要的理论和实践意义, 将为免疫学研究提供更好的解决方案, 为人类健康的改善做出贡献。

2. 模型方法

2.1. 数据预处理

我们使用的数据是已知结合的肽与 TCR 序列, 而且在数值嵌入 TCR 方面, 我们关注的是 TCR β 链的 CDR3 区域, 这是肽识别特异性的关键性决定因素。此外, 我们本次的数据预处理编码方式采用的是已经提出的自动编码器方式, 原理如下:

首先, 我们利用 Atchley factors (阿奇利因子)对氨基酸的符号进行编码, 使用 5 个数字全面表示每种氨基酸的生化特性。编码后得到的数字矩阵具有与 Atchley factors 数量相同的行数, 列数为 80。接着, 分别将肽和 TCR 序列的“Atchley 矩阵”输入到一个堆叠式自动编码器中, 这种算法能够以无监督的方式学习复杂信号。它能通过无监督的分解 - 重构过程来捕捉复杂输入的关键特征, 并以简短数字向量的形式嵌入所捕捉的输入特征。

Atchley 矩阵经过 30 个 5×2 内核的 2D 卷积层处理, 接着使用“SELU”函数激活, 然后经过批处理归一化层和一个 4×1 核的 2D 平均池化层。随后是另一个 2D 卷积层, 具有 20 个 4×2 内核, 以及相同的批处理归一化层和 2D 平均池化层。在池化后, 矩阵被转换为一个扁平层, 然后经过一个包含 30 个神经元的密集层, 使用“SELU”函数激活, 紧接着是 dropout rare 为 0.01 的 dropout 层, 以及另一个包含 30 个神经元的密集层, 这是自动编码器模型的瓶颈层。在瓶颈层之前的层被反转, 构成模型的解码器部分。编码器的输入与解码器的输出完全一致, 即 Atchley 矩阵。训练过程指导自动编码器重建输入数据, 并利用简单的数字向量捕捉其固有结构。训练完成后, 在自动编码器的最小全连接层(瓶颈)中形成原始序列的 30 个神经元数字向量嵌入。

我们通过比较输入的肽与 TCR 和重构的肽与 TCR, 由图 1 和图 2 可知原始 TCR 与肽和重构的 TCR 与肽几乎相同。图 1 中显示了原始、Atchley 因子编码、重建的 TCR, 图 2 显示了原始、Atchley 因子编码、重建的 peptide, 验证了该自动编码器。所以由分析可知, 肽可以通过肽嵌入以一种高度忠实的方式重建, TCR 也可以重建, 证明了该自动编码器的成功训练进行排版。

2.2. SVM 模型

支持向量机(Support Vector Machine, SVM)是一类按监督学习(supervised learning)方式对数据进行二元分类(binary classification)的广义线性分类器(generalized linear classifier), 其决策边界是对学习样本求解的最大边距超平面(maximum-margin hyperplane)。SVM 可对线性和非线性数据进行分类。其学习方法包含构建由简至繁的模型: 线性可分支持向量机、线性支持向量机以及非线性支持向量机。当训练数据线性可分时, 通过硬间隔最大化或者通过软间隔最大化, 学习一个线性分类器, 即软间隔支持向量机; 当训练数据线性不可分时, 通过使用核技巧及软间隔最大化, 学习非线性支持向量机。

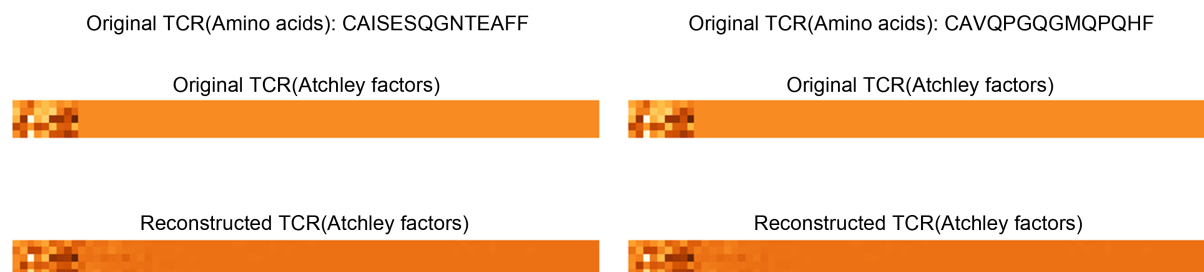


Figure 1. Comparison chart of different TCR treatments
图 1. 不同 TCR 处理后的对比图

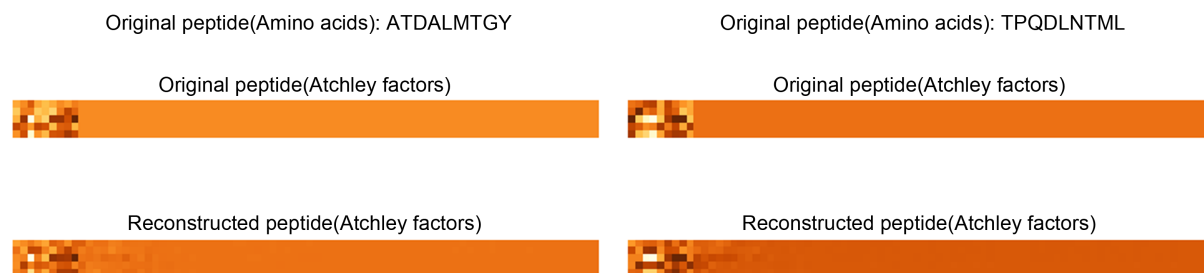


Figure 2. Comparison chart of different peptides processed
图 2. 不同 peptide 处理后的对比图

2.2.1. 软间隔支持向量机

当训练数据可以被线条分开时, 我们可以使用一个优化问题, 同时实现硬间隔最大化和软间隔最大化[3]。这个优化问题的目标是最小化一个函数, 这个函数由两部分组成: 一部分是权重向量的平方的 $1/2$ 倍, 代表了间隔的大小; 另一部分是松弛变量的加权和, 这里的 C 是一个参数, 用来控制这两部分的权衡。

约束条件是: 每个样本点必须满足, 它的标签与权重向量和这个样本点的特征向量的内积加上偏置的乘积, 至少要大于等于 1 减去松弛变量。松弛变量必须是非负的。

通过调整参数 C 的值: 当 C 非常大时, 模型会尽量减少松弛变量的值, 使得所有样本点都尽可能满足硬间隔约束, 这就是线性可分支持向量机的情况[4]。当 C 取一个有限值时, 模型会在间隔大小和允许的违反间隔的数据点数量之间进行权衡, 这就是线性支持向量机的情况。

2.2.2. 非线性支持向量机

对于非线性支持向量机, 当训练数据无法用一条直线分开时, 我们使用核技巧来处理[5]。

核技巧的思想是将输入数据映射到一个高维特征空间, 使得在这个高维空间中, 数据变得线性可分[6]。这个优化问题的目标和线性支持向量机一样, 但是约束条件是每个样本点在映射后的高维特征空间中, 它的标签与权重向量和这个样本点的高维特征向量的内积加上偏置的乘积, 至少要大于等于 1 减去

松弛变量。

2.2.3. 核函数

通过核函数的引入, 将线性不可分的数据映射到一个高维的特征空间内, 使得数据在特征空间内是可分的, 常见的核函数包括线性核函数、多项式核函数、高斯径向基函数(RBF)核函数[7]。

线性核函数(Linear Kernel): 它直接进行特征空间中的线性内积运算, 适用于线性可分的情况[8]。

多项式核函数(Polynomial Kernel): 多项式核函数通过将数据映射到高维空间并进行多项式运算来处理非线性问题[9]。

高斯径向基函数(RBF)核函数(Gaussian Radial Basis Function Kernel): RBF 核函数是最常用的核函数之一, 也称为高斯核函数, 它将数据映射到无穷维的特征空间[10]。

2.3. 集成学习

我们采用的是集成学习中的堆叠法来预测 peptide 和 TCR 的特异性结合[11], 这种方法的核心思想是利用不同基学习器的优势, 通过集成它们的预测结果来提高模型的泛化能力, 而且可以有效地捕捉基学习器之间的互补性, 从而提高整体性能。

首先将通过自动编码器的生成的肽与 TCR 的数值编码矩阵合并为特征矩阵, 再将原数据集中的 label 作为标签。其次, 将数据集分为训练集和测试集, 测试集占总数据的 20%, 并设立随机种子, 确保每次划分的结果一致。然后将训练集划分为 5 折进行交叉验证, 同时确保每个类别在每个折叠中的比例相似。接下来, 我们采用了两种方式建立模型预测 peptide 和 TCR 的特异性结合。

第一种是只建立一个简单的 SVM 模型[12], 设定核函数类型为 linear, 拟合训练数据, 最终将模型用在测试集上进行预测, 得到模型在测试集上的准确率。

第二种则是集成学习, 我们选择了三种不同内核的支持向量机(SVM)作为基学习器, 分别是线性核、径向基函数核和多项式核。对于每个基学习器, 训练集的一部分被用于训练, 另一部分被用作验证, 进行交叉验证。通过在训练集上训练的模型对验证集进行预测, 生成的预测结果构成了训练集的元特征。同时, 对测试集的预测结果进行了累加, 以便后续计算平均预测结果。另一方面, 对于元学习器的训练, 在定义了 SVC 的参数空间后, 使用网格搜索寻找最佳参数组合, 再将找到的最佳参数组合用于实例化元学习器。

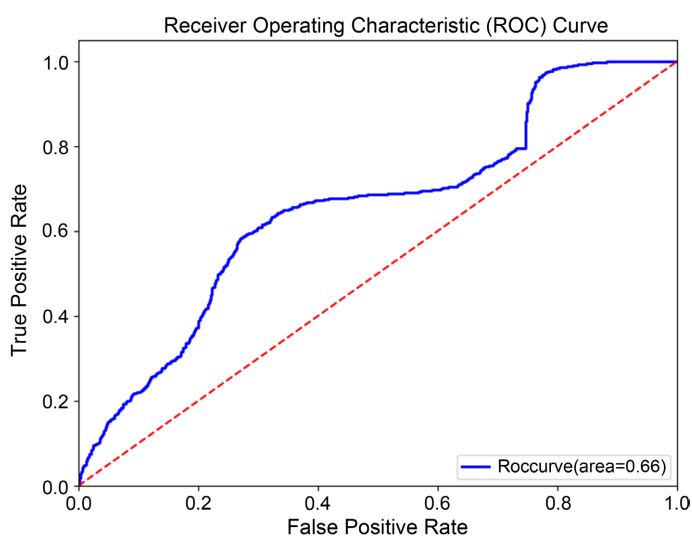


Figure 3. ROC curve graph under a single SVM model

图 3. 单一 SVM 模型下的 ROC 曲线图

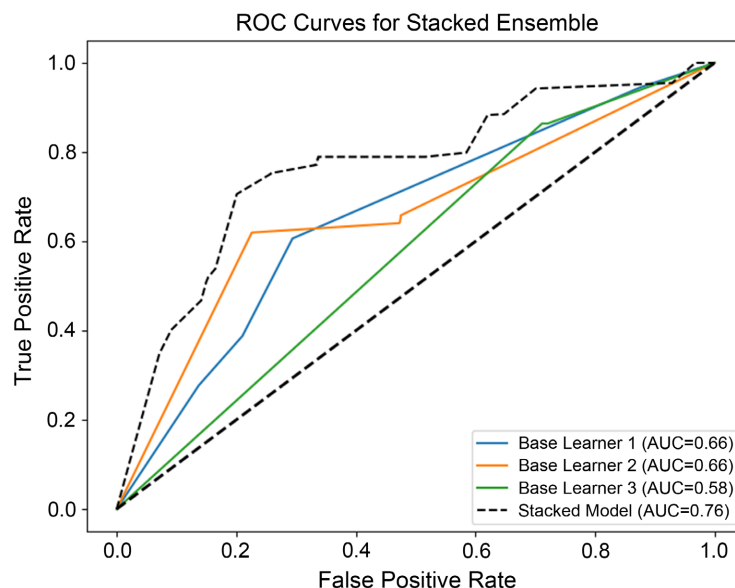


Figure 4. ROC curve graph under ensemble learning model

图 4. 集成学习模型下的 ROC 曲线图

我们令元学习器在训练集的元特征上进行训练, 学习基学习器的集成模式。最后, 使用训练好的元学习器在测试集的元特征上进行预测, 计算模型的准确率。进一步地, 我们计算每个基学习器和堆叠模型的 ROC 曲线和 ROC 值, 如上图, 图 3 中是在简单 SVM 模型下预测 peptide 和 TCR 的特异性结合的 roc 曲线, 图 4 是在集成学习模型下预测 peptide 和 TCR 的特异性结合的 roc 曲线, 根据这两种曲线以评估二分类模型的性能。

3. 结果

在单一的 svm 模型下, 即使进行了找寻最佳参数, 但最终结果并不理想, roc 只有 0.66。

而集成学习模型中, 基学习器 1 的 roc 值为 0.66, 基学习器 2 的 roc 值为 0.66, 基学习器 3 的 roc 值为 0.58, 集成之后, 得到的模型 roc 值提高到 0.76。可以看到, 集成学习模型的性能要比单一的 svm 模型性能好, 预测 peptide 和 TCR 的特异性结合的正确率高。

参考文献

- [1] 黄金海, 魏敏杰. 肿瘤抗原 CTL 表位修饰策略研究进展[J]. 现代肿瘤医学, 2016, 24(13): 2170-2174.
- [2] 王耀伟, 刘原麟, 潘若禹, 陈宝文, 罗方欣. 基于支持向量机的物流传送带场景下 RFID 识读性能研究[J]. 现代电子技术, 2024, 47(9): 150-156.
- [3] 谢松. 多摄像头非刚体目标检测与空间定位系统[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.
- [4] 李杰. 基于穿戴式传感器的人体动作识别算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2017.
- [5] 林榆坚, 梁宁波. 以未知对未知——智能安全自我进化[J]. 通信技术, 2018, 51(8): 1949-1955.
- [6] 吴今培. 基于核函数的主成分分析及应用[J]. 系统工程, 2005, 23(2): 117-120.
- [7] 徐谦. 支持服务生命周期战略管理的业务协同技术研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2017.
- [8] 宋欢. 基于警度组合预测 SVM 警信号识别模型及其算法研究[D]: [硕士学位论文]. 南昌: 江西师范大学, 2018.
- [9] 胡启国, 汪文珺. 核最小均方算法的特征映射和参数选择[J]. 南方农机, 2017, 48(13): 136-137.
- [10] 刘泽坤. 基于叠加自编码器的输气管道泄漏口径识别与定位研究[D]: [硕士学位论文]. 青岛: 山东科技大学, 2020.

- [11] Pasha, S.T., Sikder, S., Rahman, M.M., Islam, A., Alam, M.Z., Habib, M.T. and Amin, M.A. (2024) IDF23-0446 Using an Ensemble Machine Learning Model with Explainable AI (XAI) to Diagnose Gestational Diabetes Mellitus. *Diabetes Research and Clinical Practice*, **209**, Article ID: 111504. <https://doi.org/10.1016/j.diabres.2024.111504>
- [12] Chen, D.Y. and Zhong, S.P. (2012) A Universal Steganalysis Algorithm for JPEG Image Based on Selective SVMs Ensemble. *Advanced Materials Research*, **532-533**, 1548-1552. <https://doi.org/10.4028/www.scientific.net/AMR.532-533.1548>