文章编号:1008-1534(2019)05-0314-06

# MLEP:一种 B 细胞线性表位预测方法

羊红光[1],成 彬[1,2]

(1.河北省科学院应用数学研究所,河北石家庄 050081;2.河北省信息安全认证工程技术研究中心,河北石家庄 050081)

**摘 要**:为了更快更准地确定 B 细胞线性表位,提出了一种新的预测方法——MLEP(Prediction of epitope based on MCFS and LSTM,MLEP)算法。采用 5 种性质氨基酸理化性质作为学习特征,利用多聚类特征选择算法进行特征选择,用降维后的数据作为输入,用长短期记忆网络进行训练,获得预测性能好的模型,对多聚类特征选择算法及 MLEP 算法的性能进行评价。对非冗余 LBtope 数据集进行多组实验,结果表明,使用多聚类特征选择算法降维到 25 时获取性能最优模型,多聚类特征选择算法比主成分分析法获得的模型准确率更高,基于 MLEP 算法获得的模型准确率达到 94.81%。因此,MLEP 算法能更好地预测 B 细胞线性表位,对于表位预测研究具有一定的参考价值。

**关键词**:生物信息论与生物控制论;B 细胞;线性表位预测;长短期记忆网络;多群集;特征选择

**中图分类号**:R392.9 **文献标志码**:A **doi**: 10.7535/hbgykj.2019yx05004

# MLEP: A new method for prediction of linear B-cell epitopes

YANG Hongguang[1], CHENG Bin[1,2]

(1.Institute of Applied Mathematics, Hebei Academy of Sciences, Shijiazhuang, Hebei 050081, China; 2.Hebei Authentication Technology Engineering Research Center, Shijiazhuang, Hebei 050081, China)

**Abstract**: In order to determine the linear B-cell epitope faster and more accurately, a new prediction method MLEP algorithm is provided. Firstly, all the prediction calculations are based on the five properties scales of amino acids. Based on these results, a multi-cluster feature selection algorithm is studied for reducing the number of dimensions. Secondly, the networks is trained using long-short term memory network algorithm and with the reduced dimension data. Finally, the performance of the multi-cluster feature selection algorithm and the MLEP algorithm is evaluated. The experimental evaluation of classification is performed using the non-redundant LBtope dataset. The results show that the multi-cluster feature selection algorithm achieves the best performance when the dimension is reduced to 25, and the performance of the multi-cluster feature selection algorithm is significantly better than the methods based on the principal component analysis, and the maximum accuracy of 94.81% can be achieved using the MLEP algorithm. This method can effectively predict the linear epitope of B cells, which provides reference for the study of epitope prediction.

收稿日期:2019-06-26;修回日期:2019-08-16;责任编辑:张 军

基金项目:河北省重点研发计划国际科技合作专项(18390308D);河北省科学院两院合作项目(191404)

第一作者简介:羊红光(1979—),男,河北成安人,助理研究员,主要从事机器学习算法、抗原表位预测方面的研究。

通信作者:成 彬研究员。E-mail:1561862620@qq.com

羊红光,成彬.MLEP:一种 B 细胞线性表位预测方法 [J].河北工业科技,2019,36(5):314-319.

YANG Hongguang, CHENG Bin.MLEP: A new method for prediction of linear B-cell epitopes[J].Hebei Journal of Industrial Science and Technology,2019,36(5):314-319.

**Keywords**：bioinformatics and biocybernetic；B-cell；linear epitope prediction；long-short term memory；multi-cluster；feature selection

表位是抗原与抗体产生反应的区域，B 细胞表位的准确识别是表位疫苗设计、免疫诊断试剂盒开发的关键步骤之一。从结构上看，B 细胞表位分为线性表位和构象性表位，线性表位由蛋白一级序列中连续的氨基酸序列片段构成，构象性表位由空间结构相邻而在蛋白一级序列中离散分布的氨基酸序列片段组成[1]。

准确识别 B 细胞表位的方法有基于质谱的方法、基于结晶学的方法等，但这些方法存在实验复杂、设备昂贵、操作技术要求高等因素，是影响表位疫苗研发的重要因素。随着表位数据库的建立，基于机器学习的 B 细胞表位预测方法快速发展，已经成为一种速度快、成本低的有效方法[2-4]。

对 B 细胞线性表位预测的研究主要包括 2 个方面，一方面是多特征参数的复合及特征选择，另一方面是设计性能更强的表位预测模型[5-9]。表位预测的特征参数除了常用的氨基酸理化性质外，还有溶剂可及性、二级结构、氨基酸对等结构特点及统计学等。利用单一参数作为倾向标度的预测方案被证实性能有限，多种参数复合特征开展预测的方案逐渐显现出了优势，随着特征维度的增长计算量和计算复杂度也大幅的增加。在机器学习中，高维数的特征往往训练不出更高分类性能的模型。因此，如何合理选择特征是一个重要的问题。弓红岩[10]在特征集合中选出最优子集后获得性能更好的表位预测模型。LIU 等[11]通过主成分分析方法（principal components analysis，PCA）去掉了特征集合中无用或冗余的特征，获得具有较好性能的表位预测模型。特征选择的关键是在去掉无用、冗余特征的同时保留数据集的结构，更要保证特征集合具有更好的可区分性。多聚类特征选择（multi-cluster feature selection，MCFS）用于无监督特征选择，可以更好地保留数据的多集群结构，是一种较好的特征选择方法[12]。

基于机器学习的 B 细胞表位模型预测功能不断得到提升。LI 等[13]结合最大相关最小冗余度方法和增量特征选择方法，采用物理化学和生物化学性质、残基无序排列、序列保守性、溶剂可及性、二级结构、氨基酸在蛋白质-蛋白质界面和蛋白质表面保守的倾向、侧链碳原子数的偏差、进化过程中氨基酸的获得/损失等 8 种特征被用于编码肽，使用随机森林算法在测试数据集上分别达到了最高 63.53％的准确率。LIAN 等[14]利用多元线性回归建立了一种新的线性 B 细胞表位预测模型，在大型非冗余数据集上进行了 10 倍交叉验证测试，取得了 64.1％的准确度。SÖLLNER 等[15]将氨基酸的理化性质、邻域矩阵以及各自的概率和似然值等作为特征，每种肽的特征维数达到 1 487 个特征表示，通过结合特征选择的最近邻分类器，使用 5 倍交叉验证测试获得了 72％的准确度。WANG 等[16]比较和评价了 6 种不同的 B 细胞表位预测软件的正确预测真表位的能力，发现 Bepipred，AApred，BEST，LBtope 这 4 种预测软件表现优于随机组，最高的平均预测准确率为 79.71％。这些预测方案中都是在蛋白质一级序列中进行，却很少考虑序列中元素的相关性。长短期记忆网络（long-short term memory，LSTM）是一种用于长序列训练的方法，具有记忆机制，可将序列间的一些关联信息用于网络的学习训练中，有助于获得更高的识别准确率[17-18]。

## 1　方　法

### 1.1　数据获取

线性 B 细胞表位数据主要来自发表在重要学术期刊上、通过实验得出的表位数据，这些数据被整理后收录到 IEDB 数据库（http://www.iedb.org）中，该数据库由美国过敏与感染性疾病研究院（national institute of allergy and infectious diseases，NIAID）资助建设。在表位预测研究中，将 IEDB 数据库中收录的、已被标记为表位的肽段序列作为表位样本，再从包含表位样本的蛋白质一级序列中抽取未经标记的肽段作为非表位样本[19]。Uniport 数据库（https://www.uniprot.org/）提供蛋白质的一级序列、二级结构等很多结构信息。近年来，Abcpred，Bcpred，Chen，LBtope 等 4 个数据集[5]常被用于研究。LBtope 数据集从 IEDB 数据库中整理出 10 000 多条包含 20 个氨基酸的表位序列数据，通过去掉冗余之后形成的非冗余 LBtope 数据集有 7 824 个表位样本和 7 853 个非表位样本。本研究在 LBtope 数据集中进行训练、测试。

### 1.2 特征及选择方法

本研究以 5 种氨基酸理化性质为特征标度进行参数复合，它们分别是抗原性、亲水性、灵活性、疏水性、极性[20-21]。亲水性残基位于蛋白质表面，与抗原表位有密切的联系。极性氨基酸更容易暴露在蛋白质的外表，是判定抗原表位的一种特征依据。抗原性参数是 20 种氨基酸在抗原蛋白中出现频率的统计结果，是研究表位的一种特征参数。同样，疏水性和灵活性与表位形成相关也常用作特征参数。每个表位样本共包含 20 个氨基酸，因此每个样本的特征维数是 100。

MCFS 特征选择算法不同于其他特征选择方法，只针对每个特征独立计算的特定分数中选择排名最高的特征。MCFS 特征选择算法能保留不同特征间可能的相关性，从而产生最佳特征子集[12]。MCFS 算法包括 5 个步骤，具体如下。

输入：具有 $M$ 个特征的 $N$ 个数据点（$N$ 个数据点的集合为 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]$，$x_i\notin\mathbf{R}^M$），聚类数 $K$，所选特征的数量 $d$，最近邻居的数量 $p$，加权方案。

输出：选定 $d$ 个特征。

第 1 步：将图的顶点与数据点对应起来，对于每个数据点找到它的 $p$ 个最近邻居，在它们之间放置一条边，构造一个 $p$ 最近邻图。

第 2 步：设 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_K]$ 为包含相对于最小特征值的前 $K$ 个特征向量，求解方程 $\boldsymbol{L}y=\lambda\boldsymbol{D}y$ 中的特征向量，其中 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，$\boldsymbol{D}$ 为对角矩阵，$\boldsymbol{W}$ 为权重矩阵。

第 3 步：使用基数约束为 $d$ 的最小角回归算法（least angle regression，LARs），在条件 $|\boldsymbol{a}_k|\leqslant\gamma$（$\gamma$ 为设置的参数）下求解方程 $\min\limits_{a_k}\|y_k-\boldsymbol{X}_{a_k}^{\mathrm{T}}\|^2$，得到 $K$ 个稀疏系数向量 $\{\boldsymbol{a}_k\}_{k=1}^K\in\mathbf{R}^M$，其中 $\boldsymbol{a}_k$ 是一个 $M$ 维向量，并且 $|\boldsymbol{a}_k|=\sum_{j=1}^M|\boldsymbol{a}_{k,j}|$ 表示 $\boldsymbol{a}_k$ 的 $L1$ 范数。

第 4 步：根据方程 $\mathrm{MCFS}(j)=\max\limits_k|\boldsymbol{a}_{k,j}|$ 计算每个特征的 *MCFS* 得分。

第 5 步：根据它们的 MCFS 得分返回排名最高的 $d$ 个特征。

### 1.3 学习算法

LSTM 网络算法包含多个隐藏层，每个蕴含层可扩展多个记忆块，每个记忆块可包含多个记忆细胞。设输入为 $x_i$，记忆块在 $t-1$ 时刻的输出为 $h_{t-1}$，输入门单元、输出门单元、遗忘门单元和记忆块的偏置分别为 $b_i,b_o,b_f,b_g$，矩阵 $\boldsymbol{W}^{xg}$ 为 $x_i$ 和输入挤压单间的权值，矩阵 $\boldsymbol{W}^{hg}$ 为 $h_{t-1}$ 和输入挤压单元间的权值，矩阵 $\boldsymbol{W}^{xi}$ 为 $x_i$ 和输入门单元间的权值，矩阵 $\boldsymbol{W}^{hi}$ 为 $h_{t-1}$ 和输入单元间的权值，矩阵 $\boldsymbol{W}^{sf}$ 为 $x_i$ 和输出门单元间的权值，矩阵 $\boldsymbol{W}^{ho}$ 为 $h_{t-1}$ 和输出单元间的权值。记忆块在时刻 $t$ 的计算过程如下：

$$\begin{cases}g_t=\tanh(\boldsymbol{W}^{xg}x_t+\boldsymbol{W}^{gh}h_{t-1}+b_g),\\ i_t=\sigma(\boldsymbol{W}^{xi}x_i+\boldsymbol{W}^{hi}h_{t-1}+b_i),\\ f_t=\sigma(\boldsymbol{W}^{xf}x_i+\boldsymbol{W}^{hf}h_{t-1}+b_f),\\ s_t=s_{t-1}\cdot f_t+i_t\cdot g_t,\\ o_t=\sigma(\boldsymbol{W}^{xo}x_t+\boldsymbol{W}^{ho}h_{t-1}+b_o),\\ h_t=o_t\cdot\tanh(s_t)。\end{cases}\tag{1}$$

LSTM 网络学习算法采用反向传播优化权值和偏置。

采用 MCFS 算法进行特征选择，运用 LSTM 网络进行表位预测的算法称为 MLEP（prediction of epitope based on MCFS and LSTM）算法，其学习过程如图 1 所示。

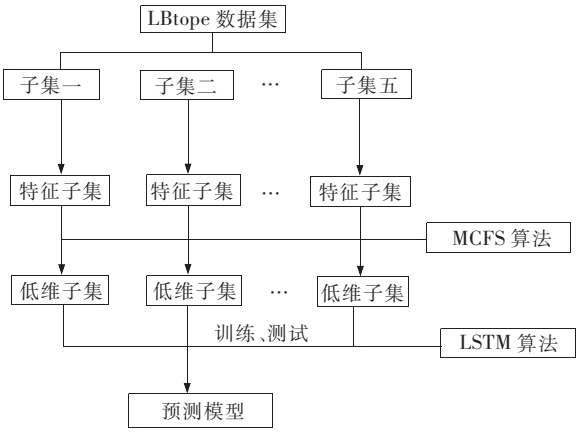


图 1 MLEP 的学习过程

Fig.1 Learning process of MLEP

### 1.4 五重验证机制

本研究将非冗余 LBtope 数据集随机分成 5 个子集（按 7 850 个数据进行划分），每个子集包含 1 570个表位样本和 1 570 个非表位样本。每次选用其中的 3 个子集作为训练集，另外 2 个子集分别作为验证和测试。一共重复 5 次这样的过程，求 5

次的平均值评价，模型性能。

### 1.5　评价参数

采用敏感性（SEN）、特异性（SPE）、阳性预测值（PPV）、准确性（ACC）、马氏相关系数（MCC）等 5 个指标评价预测模型的性能。这些指标的具体计算方法如下。

$$\mathrm{SEN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\%, \tag{2}$$

$$\mathrm{SPE}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\times 100\%, \tag{3}$$

$$\mathrm{PPV}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\%, \tag{4}$$

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\times 100\%, \tag{5}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{6}$$

式中：TP 表示正确识别的正样本（表位）数量；FN 表示模型将正样本识别为负样本（非表位）的数量；TN 表示正确识别的负样本数量；FP 表示模型将负样本识别为正样本的数量。

## 2　结果与讨论

本研究采用了五重验证机制，模型的性能指标通过 5 个子集的平均值计算获得。本研究的每个样本都包含 20 个氨基酸，采用 5 种氨基酸理化性质作为特征标度，每个特征向量的维数是 100 维。在特征集合中使用 MCFS 算法进行特征选择，可以获得指定维数的特征为输入。为了更好地验证 MLEP 算法的有效性，引入了支持向量机学习（support vector machine，SVM）算法、PCA 方法进行实验对比，共设计了 2 组实验。

### 2.1　MCFS 特征选择的有效性评价

为验证 MCFS 特征选择的有效性，将参数 $d$ 分别设为 15，25，35，45，计算后获得的特征子集，这些集合的特征向量的维数分别是 15，25，35，45。将特征子集作为输入，利用 LSTM 网络算法进行训练和测试，其性能如表 1 所示。

表 1　4 种降维情况下的性能比较

Tab.1　Performance comparison of four dimensionality reduction cases

| 维数 | SEN/% | SPE/% | PPV/% | ACC/% | MCC |
|---|---|---|---|---|---|
| 15 | 85.82 | 85.65 | 85.72 | 85.73 | 0.714 7 |
| 25 | 95.95 | 93.68 | 93.84 | 94.81 | 0.896 5 |
| 35 | 89.45 | 80.19 | 81.93 | 84.82 | 0.699 4 |
| 45 | 73.45 | 94.11 | 92.61 | 83.76 | 0.690 3 |

通过表 1 的数据，分析如下。

1）维数是 25 的模型准确率最高，达到了 94.81%，相比其他维数高出 10%左右。维数是 25 的模型的敏感性、特异性、阳性预测值 3 个指标都在 93%以上，数值相对较为均衡。这说明了 MCFS 算法能找到各类中相关性大的有用特征，将特征集合中结构较好的保留下来。

2）维数是 45 的模型敏感性值较低，而特异性、阳性预测值偏高，这说明选择这种情况下的特征对非表位样本具有一定的偏好。

3）通过 MCFS 算法选出的特征，子集维数都小于 50，训练出的模型性能较好，这与所选特征的数量小于 50 时，MCFS 算法具有好的性能表现预期一致。总的来看，基于 LSTM 学习算法利用 MCFS 算法进行特征选择后取得的表位预测模型有较高的性能。

### 2.2　MLEP 算法的性能评价

为了验证 MLEP 算法的有效性，从 2 个方面进行实验比较。1）评价 MCFS 方法是否比其他选择方法更具优势，选择了 PCA 方法与其比较。2）评价 MLEP 算法是否具有更好的性能表现，采用 PCA＋SVM，PCA ＋ LSTM，MCFS ＋ SVM 和 MLEP（MCFS＋LSTM）4 种方案进行比较。

在实验中 PCA 方法降维后的特征向量维数是 30，MCFS 方法选择 2.1 中性能最好的降维结果，即特征向量维数是 25。将降维后的特征子集作为输入，采用上述 4 种方案分别进行训练测试，实验都采用 5 重验证机制，获得的平均结果如表 2 所示。

表 2　4 种组合下的性能比较

Tab.2　Performance comparison of predictive models under four combinations

| 方法 | SEN/% | SPE/% | PPV/% | ACC/% | MCC |
|---|---|---|---|---|---|
| PCA＋SVM | 83.45 | 68.02 | 62.45 | 74.03 | 0.502 4 |
| PCA＋LSTM | 84.43 | 88.32 | 87.90 | 86.37 | 0.728 1 |
| MCFS＋SVM | 85.28 | 90.04 | 89.58 | 87.66 | 0.754 1 |
| MLEP | 95.95 | 93.68 | 93.84 | 94.81 | 0.896 5 |

通过表 2 的数据，分析如下。

1）MLEP 算法获得的预测模型准确率最高为 94.81%，从敏感性、特异性、阳性预测值等指标上看，该模型对表位、非表位均能很好的识别。

2）MCFS 算法选择特征后训练出的预测模型性能更优。使用 SVM 学习算法，PCA 方法选择特征下预测模型的准确率是 74.03%，而 MCFS 算法选择特征下预测模型的准确率是 87.66%，准确率相差 13%，这说明基于 SVM 学习算法 MCFS 选择特征下获得的预测模型性能更优。使用 LSTM 网络学习算法，PCA 方法选择特征下预测模型的准确率是 86.37%，而 MCFS 算法选择特征下预测模型的准确率是 94.81%，这说明基于 LSTM 网络学习算法 MCFS 选择特征下获得的预测模型性能更优。在两种算法下，MCFS 选择特征下获得的预测模型性能都是最优的。

3）LSTM 网络学习算法训练出的预测模型性能更优。使用 PCA 方法选择特征，LSTM 网络学习算法比 SVM 学习算法的模型准确率高 12%。使用 MCFS 算法选择特征，LSTM 网络学习算法比 SVM 学习算法的模型准确率高 7%。这说明 LSTM 网络学习算法在表位预测应用中具有一定的优势。

综合以上分析，MCFS 方法、LSTM 网络学习算法在表位预测中均有好的表现，也充分说明 MLEP 算法是一个最佳的方法，

### 2.3 讨论

线性表位预测是基于机器学习的一个分类过程，随着越来越多的特征用于学习，高维度数据处理往往需要很长的计算时间和巨大的计算开销，这也使得表位预测模型越来越难。解决这样问题的可靠方案是特征选择技术，就是在特征集合中找到相关的特征子集来降低维数。表位预测的特征提取没有固定的方案，实际研究中存在很多种组合方案，这也为特征选择带来了一定的困难。

在本研究中，尝试使用 MCFS 方法进行特征选择，一方面因为 MCFS 方法可以设定选择特征数量，具有很好的灵活性，另一方面 MCFS 方法在维数小于 50 下，能很好的将集合中的相关特征选出来，从而获得更好的预测性能。

LSTM 网络在语音识别方面具有很好表现，因为它能基于上下文中固定窗口内容对后续词进行预测。线性表位预测是基于蛋白质一级序列的，表位是序列中连续的子序列，它们之间也必然存在一定的关联关系。LSTM 网络学习算法在学习中加入记忆机制，可通过序列间的相关信息增强了学习的效果。本研究期待发挥 LSTM 网络这一优势，捕捉序列间的上下文关系实现更好的分类。实验结果表明，基于 LSTM 网络学习算法获得使表位预测模型具有更高的准确率，也明显优于其他的方法。特别地，基于 MCFS 方法和 LSTM 网络的 MLEP 算法是一个优秀的预测方法，这两者的结合进一步提高了表位的预测水平。

## 3 结 语

提出了一个新的、有效的 B 细胞线性表位预测方法——MLEP 算法，首先使用 5 种氨基酸理化性质作为特征标度，采用 MCFS 算法进行特征选择。然后，把降维后的数据作为输入，使用 LSTM 网络进行训练，获得性能优异的表位预测模型。最后，对 MCFS 算法的特征选择有效性、MLEP 算法的性能进行评价。在非冗余 LBtope 数据集进行分类实验结果说明，相比 SVM，PCA 等方法组成的方案，MLEP 算法获得最优预测模型，预测准确率达到 94.81%。

表位预测不仅需要对特征进行有效的选择，还需要更合适的学习算法训练模型。下一步工作中，将在本文基础上采用更多的特征标度，更多的特征选择方法和学习算法来评价 MLEP 算法的性能，发现具有更强性能的预测模型。

### 参考文献/References：

[1] 程华，成彬，羊红光.线性 B 细胞表位预测方法研究进展[J].中国免疫学杂志，2017，33(9)：1422-1429.

[2] 卢杨.基于蛋白质侧链信息的 B 细胞表位预测的机器学习方法[D].长春：东北师范大学，2012.
LU Yang. Machine Learning Method for B-cell Epitope Prediction Based on Protein Side Chain Information [D]. Changchun: Northeast Normal University, 2012.

[3] ELMANZALAWY Y, HONAVAR V. Recent advances in B-cell epitope prediction methods[J]. Immunome Research, 2010, 6:S2.

[4] AHMAD T, EWEIDA A, SHEWEITA S. B-cell epitope mapping for the design of vaccines and effective diagnostics[J]. Trials in Vaccinology, 2016, 5: 71-83.

[5] HARINDER S, RAHMAN A H, RAGHAVA G P S.

Improved method for linear B-cell epitope prediction using antigen′s primary sequence [J]. PLoS One, 2013, 8(5): e62216.

[6] HU Y J, LIN S C, LIN Y L, et al. A meta-learning approach for B-cell conformational epitope prediction [J]. BMC Bioinformatics, 2014, 15:378.

[7] REN Jing, LIU Qian, ELLIS J, et al. Positive-unlabeled learning for the prediction ofconformational B-cell epitopes [J]. BMC Bioinformatics, 2015, 16:S12.

[8] MOGHRAM B, NABIL E, BADR A. Ab-initio conformational epitope structure prediction using genetic algorithm and SVM for vaccine design[J].Computer Methods and Programs in Biomedicine, 2018, 153: 161-170.

[9] ZHAO Liang, WONG L, LU Lanyuan, et al. B-cell epitope prediction through a graph model[J]. BMC Bioinformatics, 2012, 13:S20.

[10] 弓红岩. 基于特征选择的线性 B 细胞表位的预测[D].大连：大连海事大学, 2018.
GONG Hongyan. Prediction of Linear B-cell Epitopes Based on Feature Selection [D]. Dalian: Dalian Maritime University, 2018.

[11] LIU Lingyun, YANG Hongguang, CHENG Bin. Prediction of linear B-cell Epitopes with PCA Method[C]// Proceedings of 2019 7th International Conference on Bioinformatics and Computational Biology. New York: USA IEEE Press, 2019: 39-43.

[12] CAI Deng, ZHANG Chiyuan, HE Xiaofei. Unsupervised feature selection for multi-cluster data[C]//Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: USA ACMPress, 2010: 333-342.

[13] LI Biqing, ZHENG Lulu, FENG Kaiyan, et al. Prediction of linear B-cell epitopes with mRMR feature selection and analysis[J]. Current Bioinformatics, 2016, 11(1): 22-31.

[14] LIAN Yao, GE Meng, PAN Xianming. EPMLR: Sequence-based linear B-cell epitope prediction method using multiple linear regression[J]. BMC Bioinformatics, 2014, 15:414.

[15] SÖLLNER J, MAYER B. Machine learning approaches for prediction of linear B-cell epitopes on proteins[J]. Journal of Molecular Recognition, 2006, 19(3):200-208.

[16] WANG Xiangyu, REN Zhonglu, SUN Qi, et al. Evaluation and comparison of newly built linear B-cell epitope prediction software from a users′ perspective [J]. Current Bioinformatics,2018, 13(2): 149-156.

[17] HUA Yuxiu, ZHAO Zhifeng, LI Rongpeng, et al. Deep learning with long short-term memory for time series prediction[J]. IEEE Communications Magazine, 2019, 57(6): 114-119.

[18] CHENG Bin, LIU Lingyun, QI Zhaohui, et al. Prediction of continuous B-cell epitopes using long short term memory networks [C]//Proceedings of 2018 6th International Conference on Bioinformatics and Computational Biology. New York: USA ACMPress,2018: 55-59.

[19] SAHA S, RAGHAVA G P. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network[J]. Proteins, 2006, 65(1): 40-48.

[20] HABIBI M, BAKHSHI P K, AGHDAM R. LRC: A new algorithm for prediction of conformational B-cell epitopesusing statistical approach and clustering method[J]. Journal of Immunological Methods, 2015,427: 51-57.

[21] DALKAS G, ROOMAN M. SEPIa, a knowledge-driven algorithm forpredicting conformational B-cell epitopesfrom the amino acid sequence[J].BMC Bioinformatics, 2017, 18:95.