

基于词嵌入的机器学习方法预测RNA柔性

朱晓锋, 常富斌, 李春华*

北京工业大学化学与生命科学学院, 北京

收稿日期: 2024年8月22日; 录用日期: 2024年9月15日; 发布日期: 2024年9月24日

摘要

RNA分子的动力学与其功能密切相关。RNA分子的柔性, 作为其动力学最基本的特性之一, 已被广泛用于研究其折叠性质、结构稳定性和配体结合能力等诸多方面。实验测定RNA柔性的方法往往比较耗时费力, 因此急需发展一种快速、准确的理论方法来预测RNA的柔性。为此, 本文提出了一种机器学习方法RNAfwe来预测RNA柔性, 该方法采用词嵌入技术提取RNA序列特征。RNAfwe与同类基于序列的RNAflex方法比较, 结果显示: 相比于使用独热编码的RNAflex (One-Hot), RNAfwe在训练和测试集上都获得了更高的皮尔逊相关系数(PCC) 0.5017和0.4704, 这表明词嵌入相较于独热编码可从RNA序列中提取与柔性更相关的特征; 相比于利用进化信息的RNAflex (PSSM), 尽管RNAfwe的性能稍差, 但前者需要知道足够的同源序列。这项工作有助于RNA动力学性质的研究, 另外为词嵌入技术广泛用于生物信息学研究提供了支持。

关键词

RNA柔性, 词嵌入, 机器学习

Word Embedding Based Machine Learning Method for RNA Flexibility Prediction

Xiaofeng Zhu, Fubin Chang, Chunhua Li*

College of Chemistry and Life Science, Beijing University of Technology, Beijing

Received: Aug. 22nd, 2024; accepted: Sep. 15th, 2024; published: Sep. 24th, 2024

Abstract

RNA molecular dynamics is closely related to their functions. The flexibility of RNA molecules, as one of the most fundamental characteristics of their dynamics, has been widely used to study their

*通讯作者。

folding properties, structural stability, ligand binding ability and so on. Experimental methods for measuring RNA flexibility are often time-consuming and labor intensive, so there is an urgent need to develop a fast and accurate theoretical method to predict RNA flexibility. To this end, we propose a machine learning method, RNAfwe, to predict RNA flexibility, which uses the word embedding technique to extract RNA sequence features. The comparison of RNAfwe with the similar sequence-based RNAflex method shows that compared with RNAflex (One-Hot), RNAfwe obtains higher Pearson correlation coefficients (PCC) of 0.5017 and 0.4704 on both training and test sets, indicating that the word embedding could extract the more related features to flexibility from RNA sequences than the one-hot encoding. Compared with RNAflex (PSSM) which uses evolutionary information, although RNAfwe has a slightly inferior performance, the former requires the knowledge of sufficient homologous sequences. This work contributes to the study of RNA dynamic properties, and provides the support for word embedding technique to be widely used in bioinformatics research.

Keywords

RNA Flexibility, Word Embedding, Machine Learning

Copyright © 2023 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

RNA (Ribonucleic acid, 核糖核酸)是一种重要的生物分子,广泛参与到生物体内的基因表达和调控过程中。RNA 分子的动力学信息对理解其生物学功能和药物设计具有重要意义。其中,柔性作为动力学研究的起点受到了人们的广泛关注。目前柔性信息已被用于预测生物分子的活性位点[1]、热稳定性[2]、结合亲和性[3]以及去折叠行为[4]等等。现在已有多种实验方法来测定生物分子动力学的信息,如核磁共振光谱[5] [6]、荧光共振能量转移[7]和 X 射线(X-ray)晶体学实验[8] [9]等。但由于这些方法的成本较高且耗时费力,所以急需发展基于计算的理论方法来预测分子的柔性。

目前,生物分子的柔性预测主要分为两大类方法:基于动力学的方法和基于机器学习的方法。其中前者包括分子动力学模拟[10] [11]和弹性网络模型[12]等方法。它们往往用于单个体系的研究,不适用于高通量预测。机器学习方法,因其计算快速且可考虑多种因素,而受到人们广泛的关注。在使用机器学习方法预测分子柔性方面,相较于蛋白质,预测 RNA 柔性的方法少之又少。目前已有四种利用机器学习的方法预测 RNA 柔性。2010 年, Tian 等人[13]建立了仅包含 13 个核糖体 RNA (rRNA)结构的数据集,并在该数据集上训练模型,目前此方法的相关程序已不可获得。2017 年, Guruge 等人[14]新建了由 142 个结构组成的数据集,该数据集包含了更多类型的 RNA。Guruge 等人分别利用基于 RNA 序列的特征——独热编码(One-Hot encoding)和位点特异性打分矩阵(Position Specific Scoring Matrix, PSSM)建立了预测模型 RNAflex (One-Hot) 和 RNAflex (PSSM),它们在测试集上获得的预测值与真实值之间的皮尔逊相关系数(Pearson Correlation Coefficient, PCC)分别为 0.4640 和 0.5028。随后, Wei 等人[15]和 Pun 等人[16]均利用了该数据集,分别开发了 RNAbval 方法和 WPHML 方法,它们在测试集上所得到的 PCC 分别为 0.6061 和 0.5822。RNAbval 方法主要利用了预测的二级结构和溶剂可及表面积等结构信息, WPHML 主要利用了加权持续同调策略来获取基于 RNA 结构的拓扑信息。由此可见,目前仅有 Guruge 等人的方法是完全基于序列信息来预测 RNA 柔性,而其它方法均使用了结构相关的信息。

本文旨在从序列中挖掘与 RNA 柔性相关的信息,从而基于序列信息预测 RNA 柔性。在机器学习方法中,除了机器学习算法外,特征对于模型的预测同样有着重大影响。在前人的工作中,主要采用了独热编码或者 PSSM 来描述序列特征,但是独热编码难以表示序列中上下文的信息,PSSM 则需要已知足够多的同源序列,因此,在本文中我们尝试使用词嵌入技术来提取序列特征。长期以来,用向量表示单词一直是解决许多自然语言处理问题的有效方法,目前利用词嵌入方法是自然语言处理领域取得突破的主要原因[17][18]。像自然语言一样,生物分子序列可以自然地表示为字母串,例如蛋白质序列由 20 种(常见氨基酸)字母组成,核酸序列由 4 种字母组成。此外,生物分子通常由具有一定功能的元素组成,这些元素可以类比为人类语言中的单词、短语和句子[19][20][21]。生物分子的序列决定结构,这意味着从信息论的角度来看,生物分子的结构信息包含在其序列中[22]。所以,将自然语言处理中的词嵌入技术用于生物信息学中,使挖掘生物序列中隐藏的结构信息成为可能。目前,词嵌入技术已在 G 蛋白偶联受体、抗菌肽、蛋白质底物特异性以及肿瘤坏死因子研究方面取得了进展[17][23][24][25]。

本文完全基于序列信息提出了预测 RNA 柔性的新方法 RNAfwe。首先获取了数据集,然后利用训练集中 RNA 序列所建立的语料库训练词嵌入模型 word2vec [26],并为每一个生物词生成了词嵌入向量。最后将得到的向量作为特征输入到支持向量机(SupportingVectorMachine, SVM)模型中,通过 5 折交叉验证调整模型参数并在独立测试集上评价模型的性能。

2. 数据与方法

2.1. 数据来源及预处理

2.1.1. 数据来源

本研究从 Pun 等人[16]的工作中获取了数据集中所有 RNA 的 PDB 号(Protein Data Bank ID),数据集中结构的筛选过程如下:首先从 PDB 数据库下载 RNA 序列长度大于 32 个碱基并且结构分辨率优于 3Å 的具有 B 因子的 X-ray 晶体结构;然后使用 CD-HIT 软件[27]进行去冗余,去除了相似性大于 80% 的 RNA 序列后得到 142 条 RNA 序列;最后将其随机拆分,其中 75% 作为训练集,25% 作为测试集,使得训练集中包含 108 条 RNA 链,测试集中包含 34 条 RNA 链。本文中,我们用核苷酸残基中的 C1 原子的 B 因子表示核苷酸的柔性。

2.1.2. 数据预处理

RNA 的种类繁多,不同类型的 RNA 结构折叠状态和柔性相差甚远,即使同一类型的 RNA 中,由于其长度的差异,不同结构的折叠状态也会造成其柔性的巨大差异。图 1 显示了训练集中的所有 RNA 的 B 因子的箱线图。从图中可以看出,不同 RNA 之间的 B 因子值分布存在较大的差异,并且在一些结构内部也存在异常的 B 因子。因此,为了使所有结构之间的 B 因子值具有可比性,首先需要去除不同结构中 B 因子的异常值并对其进行标准化处理,使得不同结构处于同一标准下进行比较。在本研究中,使用基于中值的方法[15]去除实验误差造成的异常值。然后利用以下公式对 B 因子进行标准化处理:

$$B_i = \frac{(x_i - \mu)}{\sigma}, \quad (1)$$

其中 μ 和 σ 分别是 RNA 中 B 因子的平均值和标准差。

2.2. 词嵌入向量的生成

图 2 以两条 RNA 序列为例,展示了生物词长度为 3 时生成词向量的过程。如图 2 所示,首先利用侧窗口为 n 的滑动窗口将 RNA 序列切成长度为 s 的生物词,其中 $s = 2 \times n + 1$ 。接着用所有的生物词

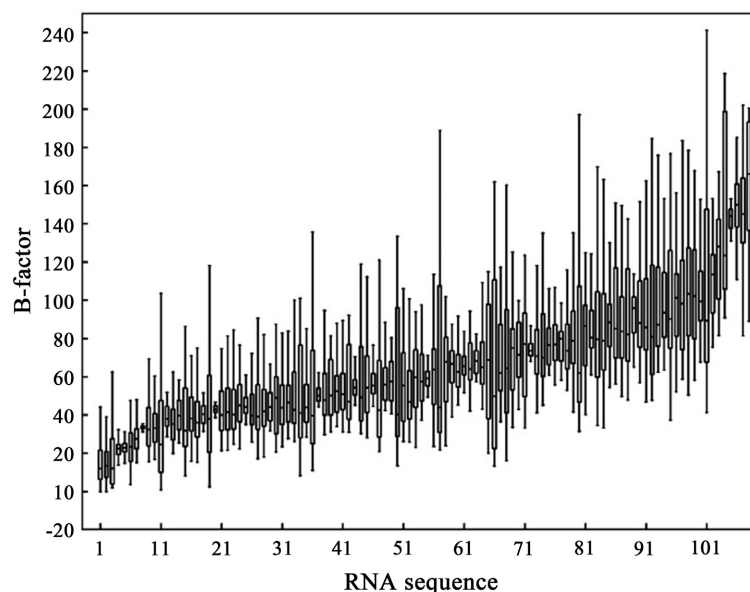


Figure 1. Boxplot of B-factors of all RNA structures in the training set

图 1. 训练集中所有 RNA 结构的 B 因子的箱线图

建立非冗余的语料库，然后用该语料库来训练 word2vec 模型并为每个生物词生成一个向量。在由语料库生成词向量的过程中，重点关注两个参数：词向量的维度 m 和上下文的生物词范围 w 。其中，词向量的维度 m 是指将一个生物词表示为 m 维的密集向量，通常来说，词向量的维度越高，其隐含的信息就越多，就可以更好的表示词语之间的关系，但维度过高会存在过拟合的风险，同时也会增加计算量，因此选择一个合适维度的词向量对于预测方法来说也是至关重要的。在 word2vec 模型中，计算机需要根据单词与其上下文之间的关系来训练出每个单词的词向量，其中 w 的大小就决定了上下文单词的范围。若 w 较大，表示目标词与序列中较远的词也会存在一定的相关性，反之则表示目标词仅与其紧邻的词相关。因此， w 的大小会较大程度影响所生成的词向量的效果。

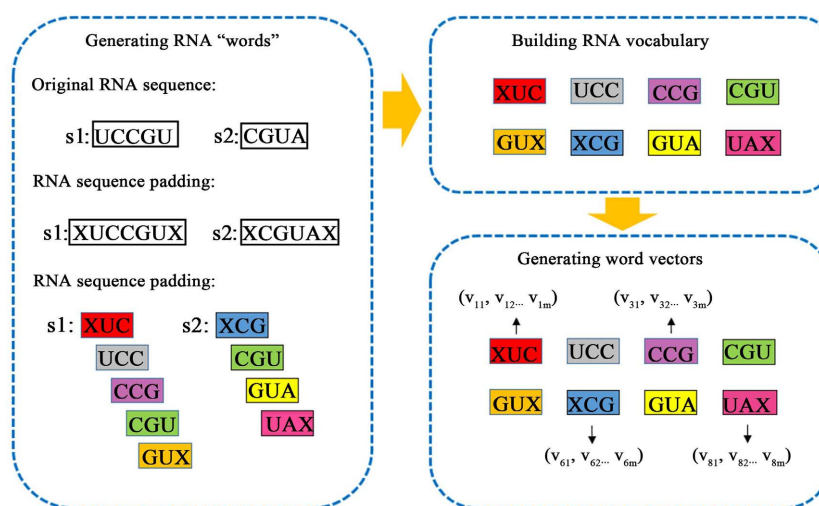


Figure 2. Generation process of the word embedding vector with 2 RNA sequences and word length = 3

图 2. 展示了 2 个 RNA 序列，生物词长度为 3 时所生成词嵌入向量的过程

2.3. 评价指标

本文中，使用皮尔逊相关系数(Pearson correlation coefficient, 简称 *PCC*)来评价模型的性能。通过计算预测 *B* 因子与真实 *B* 因子之间的 *PCC*，来测试模型的预测性能。*PCC* 计算公式为

$$PCC(B^p, B^e) = \frac{\sum_{i=1}^N (B_i^p - \bar{B}^p)(B_i^e - \bar{B}^e)}{\sqrt{\sum_{i=1}^N (B_i^p - \bar{B}^p)^2} \sqrt{\sum_{i=1}^N (B_i^e - \bar{B}^e)^2}}, \quad (2)$$

其中， B_i^p 和 B_i^e 分别表示第 *i* 个节点的预测 *B* 因子和真实 *B* 因子， $\bar{B}^p$ 和 $\bar{B}^e$ 是它们的平均值，*N* 表示样本总数。

3. 结果与讨论

3.1. 词嵌入向量参数优化

如 2.2 节所述，在生成词向量的过程中，重点考虑以下参数：生物词侧窗口长度 *n*、生成的词向量维度 *m* 和生成词向量时所考虑的上下文的范围 *w*。为了获得可以表示核苷酸的最佳词向量，采用网格搜索的方法来寻找生成词向量的最优参数组合。对于不同的参数组合，分别生成不同的词向量。本文中，*n*、*m* 和 *w* 的取值范围分别为[1、3、5、7、9]，[50、100、150]和[2、4、6、8、10]。由于 word2vec 是一个无监督的基于神经网络的词向量算法，所以在训练过程中我们通过设置迭代次数去控制其训练程度，并且通过默认参数的机器学习模型来检验所生成的词向量的效果。本文分别使用了支持向量机(Supporting Vector Machine, SVM)、随机森林(Random Forest, RF)、梯度提升树(Gradient Boosting Regression Tree, GBRT) 模型和卷积神经网络(Convolutional Neural Network, CNN)来分别寻找最优的词向量。表 1 显示了不同模型所对应的最优的词向量参数。从表 1 中我们发现这些算法所选取的最佳词向量参数非常相似，因此我们认为当 *n*、*m*、*w* 分别取 7、150 和 10 时所获得的词向量可以更好的表征核苷酸信息。此外，如表 1 所示，SVR 和 GBRT 中 *n* 均选择了较大的值(*n* = 7)，这表明核苷酸的 *B* 因子与周围较大距离内的核苷酸有关，我们认为这种情况是由 RNA 的结构较为松散，长程相互作用较多造成的。

Table 1. Best word vector parameters corresponding to different models
表 1. 不同模型所对应的最佳词向量参数

	<i>n</i>	<i>m</i>	<i>w</i>
SVR	7	150	10
RF	5	100	10
GBRT	7	150	10
CNN	7	150	10

3.2. 不同机器学习算法之间的比较

为了获得最佳的预测结果，利用 3.1 节选择的词向量作为不同模型的输入，并且使用网格搜索的方法寻找不同模型的最优超参数。经过参数搜索，不同的模型在训练集和测试集上的预测结果如表 2 所示，表中数字表示预测 *B* 因子与真实 *B* 因子之间的 *PCC*，其中最高 *PCC* 值以粗体显示。从表 2 可以看出，SVM 获得了最好的预测结果，在训练集和测试集上的 *PCC* 值分别为 0.5017 和 0.4704，优于 RF 和 GBRT。值得注意的是：卷积神经网络(CNN)在训练集上获得了与 SVM 相似的结果，但是其在测试集中略差于 SVM，我们猜测可能存在两个原因：一是我们的训练集数据量比较少，深度学习不能较好的学习其内部

的隐藏含义；二是我们可能没有找到最适合该课题的深度学习架构导致 CNN 不能充分提取序列中所隐藏的信息。

Table 2. Comparison of results from different machine learning algorithms on the training and test sets

表 2. 不同机器学习算法在训练集和测试集上的结果比较

	Training set	Test set
SVR	0.5017	0.4704
RF	0.4943	0.4246
GBRT	0.4539	0.3915
CNN	0.4987	0.4501

3.3. RNAfwe 与使用传统序列特征方法的比较

本文利用词嵌入技术提取 RNA 序列特征并结合 SVM 算法开发了从序列预测 RNA 柔性的 RNAfwe 方法，该方法在训练集和测试集上的测试结果见表 3。为了与基于传统的序列特征(独热编码和 PSSM)的方法相比，表 3 也列出了由 Guruge 等人发展的基于独热编码的方法 RNAflex (One-Hot)和基于 PSSM 的方法 RNAflex (PSSM)，在 Guruge 等人的工作中，同样使用了 SVM 算法进行预测。从表 3 可以看出：与 RNAflex (One-Hot)方法相比，RNAfwe 方法在训练集和测试集上均获得了更好的效果，分别为 0.5017 和 0.4704。我们认为这可能由于独热编码仅可表示单个核苷酸的类型，无法表示目标核苷酸的上下文序列信息且其具有高维性、稀疏性等缺点。与 RNAflex (PSSM)方法相比，RNAfwe 方法的预测结果稍差，我们认为产生这种情况的原因可能是：词向量虽然可以表示目标残基的上下文信息，但是我们无法准确得知哪些核苷酸片段具有一定的功能且无法确定具有特定功能的核苷酸片段的长度因此无法准确确定最合适的生物词长度，这就导致词向量具有不可避免的误差。然而 PSSM 可以通过多序列比对获得序列中每个部分的进化保守性，通常认为，越保守的区域其柔性就越小，所以 PSSM 可以从进化保守性层面反映更精确的柔性信息。值得注意的是，虽然 RNAflex (PSSM)得到了更好的结果，但是其需要比对大量的同源序列，在同源序列未知或数量不足的情况下基于进化的 PSSM 特征将不可用。

Table 3. Comparison between RNAfwe method and the other sequence-based methods on training and test sets

表 3. RNAfwe 与其它基于序列信息的方法在训练和测试集上的比较

	Training set	Test set
RNAfwe	0.5017	0.4704
RNAflex (One-Hot)	0.4467	0.4640
RNAflex (PSSM)	0.5176	0.5028

4. 结论

本研究探索了词嵌入技术在 RNA 柔性预测中应用，并提出了一种基于序列的 RNA 柔性预测方法 RNAfwe。该方法利用词向量表示 RNA 序列，将 RNA 中的生物词表示成一个低维的、致密的向量，然后将其作为特征输入到 SVM 模型以预测 RNA 柔性。与利用独热编码策略的方法 RNAflex (One-Hot)相比，RNAfwe 在训练集和测试集上均获得了更高的 PCC，分别为 0.5017 和 0.4704，这表明词向量相对于独热

编码更适合于 RNA 柔性预测。尽管 RNAfwe 方法的性能略次于 RNAflex (PSSM)方法, 但是后者需要目标序列存在一定数量的同源序列, 否则其无法获得准确的进化信息。

目前, 具有真实 B 因子的 RNA 结构相对较少, 这限制了利用机器学习预测 RNA 柔性方法的发展。随着实验技术的发展, 将有越来越多的 RNA 柔性数据可供使用, 这有利于训练词嵌入模型, 以获得更准确的词向量表示。在将来, 基于词嵌入的方法可以达到甚至超越基于进化信息的方法。准确预测 RNA 的柔性可以帮助人们更好地理解 RNA 的动力学机制, 揭示 RNA 结构与功能之间的关系, 并有助于设计更加稳定的 RNA 药物。因此, 这项工作对于 RNA 的相关研究具有重要意义。

基金项目

国家自然科学基金项目(32271294, 31971180)。

参考文献

- [1] Carugo, O. and Argos, P. (1998) Accessibility to Internal Cavities and Ligand Binding Sites Monitored by Protein Crystallographic Thermal Factors. *Proteins, Structure, Function, and Bioinformatics*, **31**, 201-213. [https://doi.org/10.1002/\(SICI\)1097-0134\(19980501\)31:2<201::AID-PROT9>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1097-0134(19980501)31:2<201::AID-PROT9>3.0.CO;2-O)
- [2] Schneider, B., Gelly, J., de Brevern, A.G., et al. (2014) Local Dynamics of Proteins and DNA Evaluated from Crystallographic B Factors. *Acta Crystallographica Section D Biological Crystallography*, **70**, 2413-2419. <https://doi.org/10.1107/S1399004714014631>
- [3] Liu, Q., Kwok, C.K. and Li, J. (2013) Binding Affinity Prediction for Protein-Ligand Complexes Based on β Contacts and B Factor. *Journal of Chemical Information and Modeling*, **53**, 3076-3085. <https://doi.org/10.1021/ci400450h>
- [4] Li, C., Lv, D., Zhang, L., et al. (2016) Approach to the Unfolding and Folding Dynamics of Add A-Riboswitch upon Adenine Dissociation Using a Coarse-Grained Elastic Network Model. *The Journal of Chemical Physics*, **145**, Article ID: 014104. <https://doi.org/10.1063/1.4954992>
- [5] Hu, Y., Cheng, K., He, L., et al. (2021) NMR-Based Methods for Protein Analysis. *Analytical Chemistry*, **93**, 1866-1879. <https://doi.org/10.1021/acs.analchem.0c03830>
- [6] Ishima, R. and Torchia, D. (2000) Protein Dynamics from NMR. *Nature Structural Biology*, **7**, 740-743. <https://doi.org/10.1038/78963>
- [7] Sasmal, D.K., Pulido, L.E., Kasal, S., et al. (2016) Single-Molecule Fluorescence Resonance Energy Transfer in Molecular Biology. *Nanoscale*, **8**, 19928-19944. <https://doi.org/10.1039/C6NR06794H>
- [8] Hoshino, M., Adachi, S. and Koshihara, S. (2015) Crystal Structure Analysis of Molecular Dynamics Using Synchrotron X-Rays. *CrystEngComm*, **17**, 8786-8795. <https://doi.org/10.1039/C5CE01128K>
- [9] Christoforides, E., Fourtaka, K., Andreou, A., et al. (2020) X-Ray Crystallography and Molecular Dynamics Studies of the Inclusion Complexes of Geraniol in β -Cyclodextrin, Heptakis (2, 6-di-O-methyl)- β -Cyclodextrin and Heptakis (2, 3, 6-tri-O-methyl)- β -Cyclodextrin. *Journal of Molecular Structure*, **1202**, Article ID: 127350. <https://doi.org/10.1016/j.molstruc.2019.127350>
- [10] Scott, A.H. and Ron, O.D. (2018) Molecular Dynamics Simulation for All. *Neuron*, **99**, 1129-1143. <https://doi.org/10.1016/j.neuron.2018.08.011>
- [11] McCammon, J.A., Gelin, B.R. and Karplus, M. (1977) Dynamics of Folded Proteins. *Nature*, **267**, 585-590. <https://doi.org/10.1038/267585a0>
- [12] Bahar, I., Atilgan, A.R. and Erman, B. (1997) Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential. *Folding and Design*, **2**, 173-181. [https://doi.org/10.1016/S1359-0278\(97\)00024-2](https://doi.org/10.1016/S1359-0278(97)00024-2)
- [13] Tian, F., Zhang, C., Fan, X., et al. (2010) Predicting the Flexibility Profile of Ribosomal RNAs. *Molecular Informatics*, **29**, 707-715. <https://doi.org/10.1002/minf.201000092>
- [14] Guruge, I., Taherzadeh, G., Zhan, J., et al. (2018) B-Factor Profile Prediction for RNA Flexibility Using Support Vector Machines. *Journal of Computational Chemistry*, **39**, 407-411. <https://doi.org/10.1002/jcc.25124>
- [15] Wei, H., Wang, B., Yang, J., et al. (2021) RNA Flexibility Prediction with Sequence Profile and Predicted Solvent Accessibility. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **18**, 2017-2022. <https://doi.org/10.1109/TCBB.2019.2956496>
- [16] Pun, C.S., Yong, B.Y.S. and Xia, K. (2020) Weighted-Persistent-Homology-Based Machine Learning for RNA Flexibility Analysis. *PLOS ONE*, **15**, e237747. <https://doi.org/10.1371/journal.pone.0237747>

-
- [17] Nguyen, T., Le, N., Ho, Q., *et al.* (2019) Using Word Embedding Technique to Efficiently Represent Protein Sequences for Identifying Substrate Specificities of Transporters. *Analytical Biochemistry*, **577**, 73-81. <https://doi.org/10.1016/j.ab.2019.04.011>
- [18] Goth, G. (2016) Deep or Shallow, NLP Is Breaking Out. *Communications of the ACM*, **59**, 13-16.
- [19] Solan, Z., Horn, D., Rupp, E., *et al.* (2005) Unsupervised Learning of Natural Languages. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 11629-11634. <https://doi.org/10.1073/pnas.0409746102>
- [20] Strait, B.J. and Dewey, T.G. (1996) The Shannon Information Entropy of Protein Sequences. *Biophysical Journal*, **71**, 148-155. [https://doi.org/10.1016/S0006-3495\(96\)79210-X](https://doi.org/10.1016/S0006-3495(96)79210-X)
- [21] Yu, L., Tanwar, D.K., Penha, E.D.S., *et al.* (2019) Grammar of Protein Domain Architectures. *Proceedings of the National Academy of Sciences*, **116**, 3636-3645. <https://doi.org/10.1073/pnas.1814684116>
- [22] Ptitsyn, O.B. (1991) How Does Protein Synthesis Give Rise to the 3D-Structure? *FEBS Letters*, **285**, 176-181. [https://doi.org/10.1016/0014-5793\(91\)80799-9](https://doi.org/10.1016/0014-5793(91)80799-9)
- [23] Qiu, W., Lv, Z., Xiao, X., *et al.* (2021) EMCBOW-GPCR: A Method for Identifying G-Protein Coupled Receptors Based on Word Embedding and Wordbooks. *Computational and Structural Biotechnology Journal*, **19**, 4961-4969. <https://doi.org/10.1016/j.csbj.2021.08.044>
- [24] Hamid, M. and Friedberg, I. (2019) Identifying Antimicrobial Peptides Using Word Embedding with Deep Recurrent Neural Networks. *Bioinformatics*, **35**, 2009-2016. <https://doi.org/10.1093/bioinformatics/bty937>
- [25] Nguyen, T., Le, N., Ho, Q., *et al.* (2020) TNFPred: Identifying Tumor Necrosis Factors Using Hybrid Features Based on Word Embeddings. *BMC Medical Genomics*, **13**, Article No. 155. <https://doi.org/10.1186/s12920-020-00779-w>
- [26] Tomas, M., Kai, C., Greg, C., *et al.* (2013) Efficient Estimation of Word Representations in Vector Space. *CoRR*. arXiv preprint arXiv:1301.3781
- [27] Li, W. and Godzik, A. (2006) Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics*, **22**, 1658-1659. <https://doi.org/10.1093/bioinformatics/btl158>