Hans Journal of Biomedicine 生物医学, 2021, 11(3), 121-128
Published Online July 2021 in Hans. http://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2021.113016

Hans汉斯

DNAVec：基因组DNA序列的预训练词向量表示

郎　梅，郭志云

西南交通大学生命科学与工程学院，四川 成都
Email: zhiyunguo@swjtu.edu.cn

收稿日期：2021年4月23日；录用日期：2021年5月7日；发布日期：2021年6月1日

摘　要

破译DNA序列所代表的信息是基因组研究的基本问题之一。基因调控编码由于存在多义性关系而变得非常复杂，而以往的生物信息学方法往往无法捕捉到DNA序列的隐含信息，尤其是在数据匮乏的情况下。因而从序列信息中预测DNA序列的结构和功能是计算生物学的一个重要挑战。为了应对这一挑战，我们引入了一种新的方法，通过使用自然语言处理领域的语言模型BERT将DNA序列表示为连续词向量。通过对DNA序列进行建模，BERT有效地从未标记的大数据中捕捉到了DNA序列中的序列特性。我们将DNA序列的这种新的嵌入表示称为DNAVec (DNA-to-Vector)。此外，我们可以从模型中提取出预训练的词向量用于表示DNA序列，用于其他序列级别的分类任务。

关键词

BERT，DNA序列，预训练，自然语言处理

DNAVec: Pre-Trained Word Vector Representation of Genomic DNA Sequences

Mei Lang, Zhiyun Guo

School of Life Science and Engineering, Southwest Jiaotong University, Chengdu Sichuan
Email: zhiyunguo@swjtu.edu.cn

Received: Apr. 23rd, 2021; accepted: May 7th, 2021; published: Jun. 1st, 2021

Abstract

Deciphering the information represented by DNA sequences is one of the fundamental problems

文章引用: 郎梅, 郭志云. DNAVec: 基因组 DNA 序列的预训练词向量表示[J]. 生物医学, 2021, 11(3): 121-128.
DOI: 10.12677/hjbm.2021.113016

of genomic research. Gene regulatory coding is complicated by the presence of polysense relationships, and previous bioinformatics methods often fail to capture the implicit information of DNA sequences, especially when data are scarce. Predicting the structure and function of DNA sequences from sequence information is thus an important challenge in computational biology. To address this challenge, we introduce a new approach to represent DNA sequences as continuous word vectors by using the language model BERT from the field of natural language processing. By modelling DNA sequences, BERT effectively captures the sequence properties in DNA sequences from unlabelled big data. We refer to this new embedding representation of DNA sequences as DNAVec (DNA-to-Vector). In addition, we can extract pre-trained word vectors from the model for representing DNA sequences for other sequence-level classification tasks.

Keywords

BERT, DNA Sequence, Pre-Training, Nature Language Processing

1. 引言

最近，研究人员发现人类语言与生物语言之间存在一些相似之处。受此启发，开发生物语言处理模型可以为生物序列的研究提供一个全新的理论视角和可行的方法，然而，人们要发现与人类语言不同的生物模式的真正含义是相当困难的。研究人员开发了一个叫的 BioVec [1]模型。该模型提供了一种全新的蛋白质序列的表示和特征提取方法。Wang 等人开发了一种名为生物向量(Bio2Vec)的生物语言处理模型[2]。Bio2Vec 提供了一个框架，允许研究人员考虑生物序列的上下文信息和隐含语义信息。Chen 等人[2]利用自然语言处理(NLP)技术获得了各蛋白质序列的全局向量表示。Heinzinger 等[3]利用深度双向语言模型 ELMo [4]从未标记的大数据中捕捉蛋白质序列的生物物理特性。MENEGAUX 等人[5]开发了一称之为 fastDNA 的模型，此模型通过学习 DNA 序列所包含的 k-mer 的连续低维表示，将 DNA 序列嵌入到一个向量空间中。他们通过修改 fastText 开源库[6] [7]实现了这个模型，其中涉及到一个类似的自然语言的 k-mer 嵌入模型。Patrick Ng 开发了一个名为 dna2vec [8]的方法，该方法基于 NLP 的一个开创性的模型 Word2Vec [9]来提取 DNA 序列的特征并给出适当的表示，可以更好地理解 DNA 序列的语义。以上实验结果表明，利用生物序列的语义信息对解决基于序列的问题有很大帮助。此外，这些工作将在各种生物分类问题上也有潜在的应用。

上面研究的表明 NLP 在生物语言处理领域有很大的潜力，但是自然语言处理社区技术也不断的更新和发展，以前技术的不足也不能满足当下发展的需求。比如：之前的模型 Word2Vec [9]、GloVe [10]专注于学习与语境无关的词语表征，最近的研究 ELMo [4]使用了一个双向的语言模型学习上下文表示，集中在学习依赖语境的单词表征上，ELMo 使用两个独立的长短期记忆网络(LSTM) [11]的组合，并不是真正意义上的上下文相关表示。GloVe [10]使用机器翻译将上下文信息嵌入到单词中。这些自然语言处理技术虽然得到了广泛的应用，但它们对序列的表示都不是真正意义上的上下文相关。BERT [12]是广泛用于 NLP 领域的基于上下文相关的词表示模型，同时在大多数的 NLP 应用中取得了最先进的性能。在大量自然语言处理领域的实践表明 BERT 具有很强的表示能力，而表示学习[13]对于深度学习模型是非常重要的。基于 BERT 这种对序列的表示能力，因此在这篇文章中我们基于全基因组 DNA 序列从头训练 BERT

模型，从而得到一个 DNA 序列的表示模型。模型的可视化结果表明，模型能够捕捉到 DNA 序列中特殊的语义模式。

2. 预训练模型

2.1. 数据集

我们的预训练模型是用 hg38 人类基因组组装 chr1 至 chr22 进行训练的。具体来说，它们是从 UCSC (http://hgdownload.cse.ucsc.edu/downloads.html#human)下载的。此外，我们排除了 X 和 Y 染色体，以及线粒体和未定位的序列。

2.2. BERT：基于 Transformer 的双向语言模型

BERT 是一个上下文相关的的词表示模型。该模型是基于遮蔽语言模型和预训练的使用双向 Tranformers [14]。由于语言模型的本质是未来的词不能被看见，以前的语言模型仅限于两个单向语言模型的组合(即从左到右和从右到左)。BERT 使用了一个遮蔽语言模型，可以预测序列中随机遮蔽的词，因此可以用于学习上下文相关表示。同时，它在大多数 NLP 任务上获得了最先进的性能，同时只需要最小的特定任务架构修改就能用于其他任务。根据 BERT 的作者，在自然语言模型中加入双向表征的信息，而不是单向表征的信息，对于表征自然语言中的词是至关重要的。我们假设这种双向表示对于 DNA 序列的表示也是至关重要的。因为复杂的 DNA 序列表示也不是简单的从左到右或者从右到左的关系，基因调控元件之间也是上下文相关的。

2.3. 分词

我们没有将每个碱基视为一个单一的标记，而是用 k-mer 表示法将一个 DNA 序列标记化，这种方法已被广泛用于分析 DNA 序列。k-mer 表示法通过将每个脱氧核苷酸碱基与它的后续碱基连接起来，为其整合了更丰富的上下文信息。它们的连接称为 k-mer。本文中我们使用可变长 k-mer ($3 \le k \le 8$)来组装 DNA 序列。具体方法如下：给定一个 DNA 序列 S，首先通过在 S 上滑动长度为 k 的窗口将其转换为重叠的固定长度的 k-mer，其中 k 值的选择采用离散随机采样。例如，GATCCCAC 的变长 k-mer (k = (4, 5, 6)可以是{GATC, ATCCC,TCCCAC}。分词示例如图 1 所示。在我们的实验中，模型的词汇表包括 k-mer 的所有排列组合以及 5 个特殊标记。[CLS]代表分类标记；[PAD]代表填充标记；[UNK]代表未知标记；[SEP]代表序列上下句分离标记；[MASK]代表屏蔽标记。因此模型中的词汇共有 87365 个。

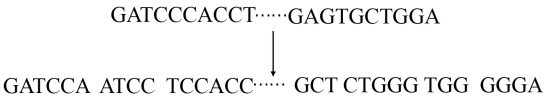

Figure 1. Variable length k-mer assembled
图 1. 可变长 k-mer 从头组装

2.4. 预训练

根据之前的模型预训练工作[12] [15]，本实验的总体结构如图 2 所示，对于一段 DNA 序列(序列的最大输入长度为 512)，我们将其标记为 k-mers 序列，并在其开头添加一个代表整个序列的特殊的标记[CLS]以及在结尾添加一个表示序列结束的特殊的标记[SEP]。在训练过程中，我们遮蔽序列中的某些 k-mers，遮蔽的比例为占一段输入序列的 15% (防止过拟合以及减少模型的计算量)。在本研究中我们使用 BERT (L = 12, H = 512, A = 12)相同的模型结构进行训练，其中 L 代表代表模型的总的层数，即有 12 个

Transformer 结构单元，H 代表隐藏层大小，A 代表自注意力头部，共有 12 个注意力头部。我们对预训练模型进行了共 80 k 步的训练，批处理量为 8。学习率为 $4e^{-4}$。此外，我们在配备 2 个 NVIDIA Tesla K80 (240 k) GPU 的机器上进行训练。

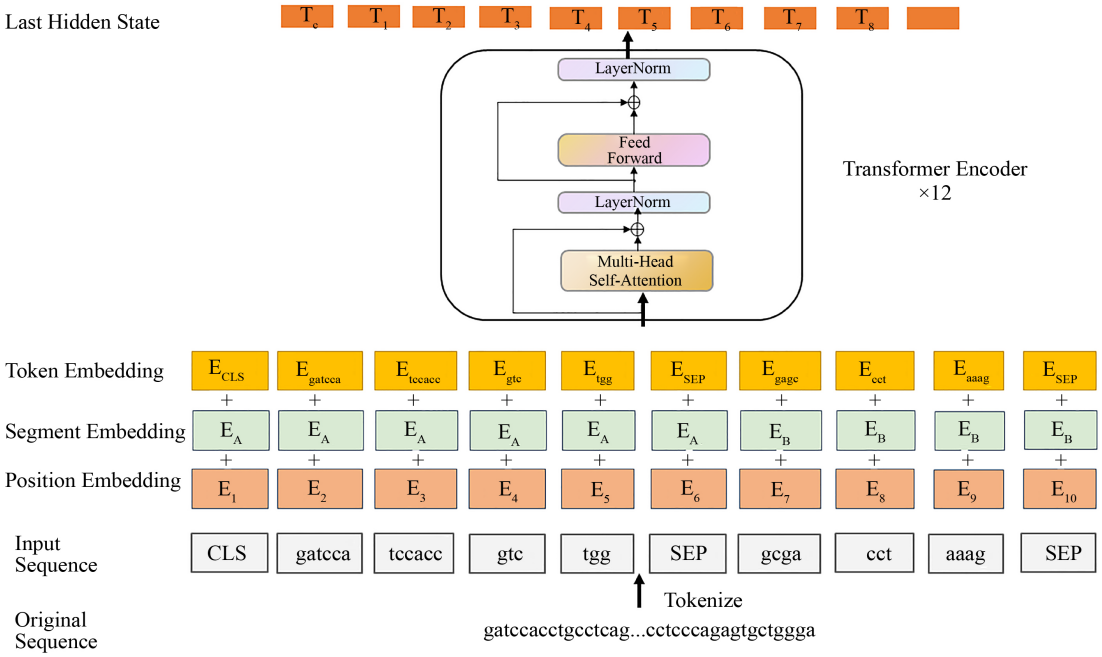

Figure 2. Pre-training model structure
图 2. 预训练模型结构

2.5. 词频－逆文档频率

tf-idf (term frequency-inverse document frequency)表示词频－逆文档频率，tf (term frequency)是词频，idf (inverse document frequency)表示逆文档频率，tf-idf 权重是信息检索和文本挖掘中常用的加权技术。该权重是一种统计度量方法，用于评估一个单词对集合或语料库中的文档的重要性。重要性随单词在文档中出现的次数成比例增加，但会被单词在包含该单词的语料库中出现的频率抵消，这有助于根据某些单词在一般情况下出现更频繁的事实进行调整。tf-idf 计算方法为：

我们定义 $f(t,d)$ 为单词 t 在文档 d 中的出现频率，然后词频 $f(t,d)$ 如公式 2-1 所示：

$$f(t,d)=\frac{f_{t,d}}{\sum_{t'\in d} f_{t',d}} \tag{2-1}$$

逆文档频率 idf 定义如公式 2-2 所示：

$$idf(t)=\log\frac{1+n}{1+df(t)}+1 \tag{2-2}$$

其中 n 在文档集中文档的总数，$df(t)$ 是包含 t 的文档的数量，tf-idf 定义如公式 2-3 所示：

$$tf\text{-}idf(t,d)=tf(t,d)\times idf(t) \tag{2-3}$$

然后将得到的 tf-idf 向量用欧几里得范数进行归一化处理，如公式 2-4 所示：

$$v_{norm} = \frac{v}{\|v\|^2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}} \tag{2-4}$$

3. 实验结果

3.1. 模型注意力机制分析

BERT 以两个关键思想为基础：Transformer [14]架构和无监督的预训练。这两个思想是 BERT 在很多应用上取的很好的原因。Transformer 是一个序列模型，它放弃了循环神经网络(RNN) [16]的循环结构，而采用了完全基于注意力的方法[17]。为了探索预训练到底学到了什么。我们使用可视化工具 bertviz [18]对模型进行可视化，通过分析模型的注意力机制我们发现一些非常独特的和令人惊讶的直观的注意力模式。下面我们发现了 4 种比较重要的注意力机制模式，并为每一个特定的 layer/head 进行分析。

模式一：注意力大部分放在 3-mer 上。在这个模式中，在特定位置的大部分注意力指向序列中的 3-mer 标记。我们可以在 layer 0/head 3 中看到这样的一个例子。(选中的头部由顶部颜色条中突出显示的正方形表示)，图 3(a)显示了对一个选定标记 mer“gatccac”的注意力模式。在本例中，几乎所有的注意力都指向序列中的下一个令牌“cct”、“aaa”、“cct”。一个词注意放在 3-mer，可能是因为 3-mer 是密码子的，在一个序列中比较重要的序列单元的原因。

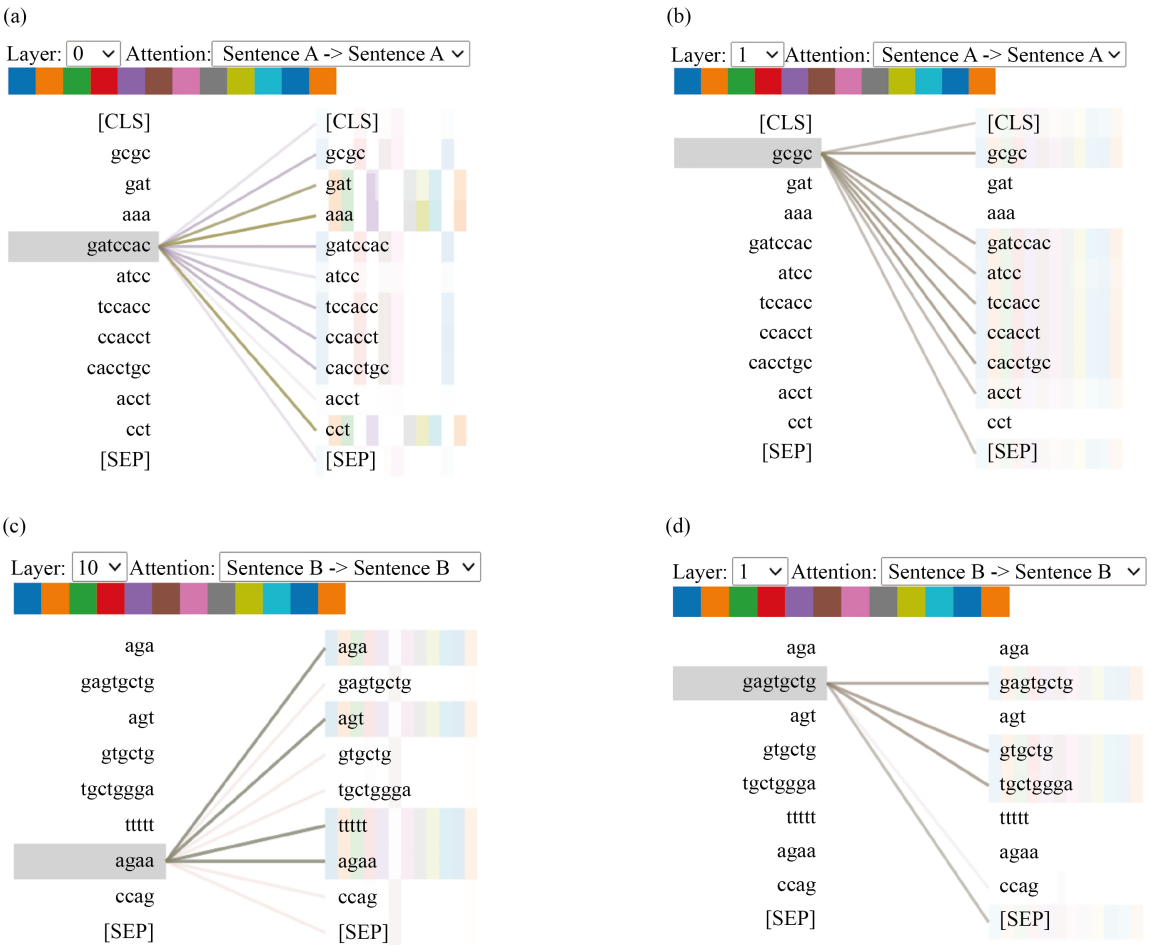

Figure 3. Visualisation of attentional mechanisms in pre-trained models
图 3. 预训练模型中的注意力机制可视化

模式二：在这个模式中，序列是一段管家基因序列。注意力放在了除 3-mer 之外的所有标记。例如，在图 3(b)中 layer3，head3 中，“gcgc”注意放在任何不包含“gat”、“aaa”、“cct”的 k-mer。一项研究用 k-mer 频率作为特征来预测 CpG 岛，结果表明人类的 CpG 岛序列具有独特的 k-mer 模式，并不是随机序列，通过他们的分析，4-mer 的表现最好[19]。表明 CpG 岛具有明显的 4-mer 分布。正如我们所知道的，CpG 岛富集于管家基因，管家基因因为要维持在所有细胞中稳定表达，而 3-mer 是比较常见的转录密码子，在基因组属于转录比较活跃的部分。因此，模式二可能学习到序列中隐含的调控语义。

模式三：注意力放在了语料库中 tf-idf 值比较高的 k-mer 上。我们计算了全基因组 DNA 序列语料库中的 k-mer 的 tf-idf 值。其中 tf-idf 值前 100 的 k-mer 的词云图如图 4 所示。根据图 3(c)可知，注意力机制放在了 tf-idf 值比较高的 mers 上。模式三表明该模型能捕捉到语料库中的关键 k-mer。

模式四：在这一模式中。注意放在了相同或相关的词，包括源词本身。在图 3(d)的例子中，“gagtgagt”的注意力主要集中在它本身和“gtgctg”、“tgctgga”上。这种模式不像其他模式那样明显，注意力分散在许多不同的单词上，但是其注意力类似于自然语言领域把注意力放在近义词上，表明模型能够捕获相似的语义。以上 4 种模式描述了预训练模型中比较常见的注意模式，结果表明该模型实际上得到了一些有意义的 DNA 序列特性。

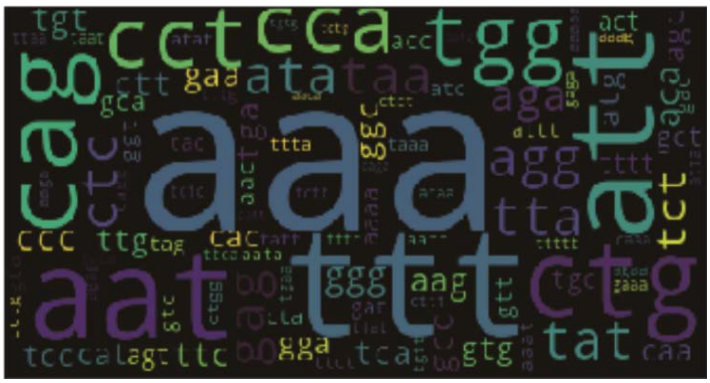

Figure 4. Word cloud graph of the top 100 kmer with relatively high tf-idf values
图 4. 前一百个 tf-idf 值比较高的 kmer 的词云图

3.2. DNA 序列的词向量表示

长 DNA 序列的普遍表示之一是将其分解为较短的 k-mer 成分，在许多应用中，短 k-mer 被认为是相关的，如 k = 6 [2]，k ≤ 7 [20]，k = 8 [21]。不幸的是，将 k-mer 作为一个 one-hot 向量进行直接的编码时任何一对 one-hot 之间的距离都是等距的，这表示 k-mer 之间完全没有相关性。而且这种编码常常导致数据的空间维度很大，容易受到维数诅咒的影响。当应用最新的机器学习算法来解决生物序列分析的问题时，这尤其成问题。因此，我们提出了一种基于预训练的 DNA 序列表示方法。考虑到 BERT 用于预训练的字典不能太大，我们使用长度为 3~8 的 k-mer。训练模型将每个 k-mer 嵌入一个新的 n 维特征空间，为不同长度的 k-mer 生成特征向量。具体地，预训练嵌入模型可以表示为大小为$V \times N$ 的投影矩阵，其中 V 为词典大小，N 为嵌入特征空间的维数，词汇表 V 是所有 k-mer 的组合和用于标记的 5 个特殊符号的集合。在本论文中，我们提取预训练模型的最后一层隐含层，得到一个大小为$V \times 512$ 的嵌入矩阵。每个单词都嵌入在一个 512 维的空间中。我们可以使用一个名为 bert-as-service 的工具(https://github.com/hanxiao/bert-as-service)获取上下文单词嵌入。获得的预训练的词向量形式如图 5 所示。

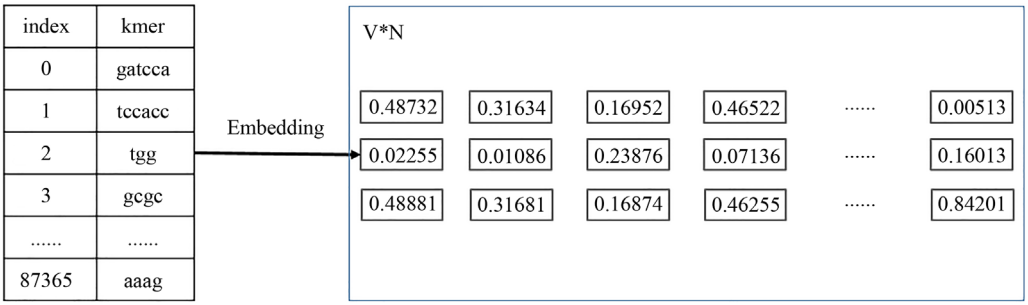

Figure 5. Pre-trained word vector representation
图 5. 预训练词向量表示形式

4. 结论

目前预训练技术广泛应用于自然语言处理(NLP)和计算机视觉(CV)领域。但是在生物信息领域还很少应用到。DAN 序列作为生物语言，其本身和自然语言有一定的相似性。因此本论文试图从语言处理的角度分析生物语言。希望为 DNA 序列的解读提供新的发现或为生物问题的解决提供新的视角。在本研究中，通过自监督的预训练方法来得到 DNA 序列的上下文相关表示，这与之前的序列 one-hot 方法表示或者序列上下文无关表示极为不同。论文的主要贡献是发布了一个预训练模型，我们期望我们的模型也适用于其他序列分析任务，例如，从染色质可及性测序数据[22]和转录因子结合位点测序数据[23]中确定基因组调控元件。此外，由于 RNA 序列与 DNA 序列只相差一个碱基，而语法和语义基本保持一致，我们提出的方法预计也可能应用 RNA 序列数据[24]。虽然在 DNA 上像自然语言样直接进行机器翻译是不可能的，但 DNA 序列的预训练模型提供了这种可能性的启示。作为一个基于 DNA 序列开发的预训练语言模型，它正确地捕捉到了 DNA 序列中隐藏的语法和语义。同时，本论文也凸显了结合不同层次的数据对 DNA 序列进行解读的必要性。综上所述，我们预计此预训练模型可以为基因序列分析带来先进的语言建模视角，为生物信息学界带来新的见解。

参考文献

[1] Asgari, E. and Mofrad, M.R. (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS ONE*, **10**, e0141287. https://doi.org/10.1371/journal.pone.0141287

[2] Chen, Y.H., Nyeo, S.L. and Yeh, C.Y. (2005) Model for the Distributions of k-Mers in DNA Sequences. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, **72**, Article ID: 011908. https://doi.org/10.1103/PhysRevE.72.011908

[3] Heinzinger, M., *et al*. (2019) Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences. *BMC Bioinformatics*, **20**, 723. https://doi.org/10.1186/s12859-019-3220-8

[4] Peters, M.E., *et al*. (2018) Deep Contextualized Word Representations.

[5] Menegaux, R. and Vert, J.P. (2019) Continuous Embeddings of DNA Sequencing Reads and Application to Metagenomics. *Journal of Computational Biology*, **26**, 509-518. https://doi.org/10.1089/cmb.2018.0174

[6] Joulin, A., *et al*. (2017) FastText.zip: Compressing Text Classification Models.

[7] Joulin, A., *et al*. (2016) Bag of Tricks for Efficient Text Classification. *Proceedings of the* 15*th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, 427-431. https://doi.org/10.18653/v1/E17-2068

[8] Ng, P. (2017) dna2vec: Consistent Vector Representations of Variable-Length k-Mers.

[9] Mikolov, T., Sutskever, I. and Chen, K. (2013) Distributed Representations of Words and Phrases and Their Compositionality.

[10] Pennington, J., Socher, R. and Manning, C.D. (2015) GloVe: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1532-1543.

https://doi.org/10.3115/v1/D14-1162

[11] Greff, K., Koutnık, R.K.S.J. and Schmidhuber, B.R.S.J.U. (2015) LSTM: A Search Space Odyssey.

[12] Devlin, J., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[13] Bengio, Y., Courville, A. and Vincent, P. (2013) Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

[14] Shazeer, N., *et al*. (2017) Attention Is All You Need.

[15] Diederich, A. (2019) Advances in Neural Information Processing Systems 18. *Journal of Mathematical Psychology*, **51**, 339. https://doi.org/10.1016/j.jmp.2008.09.003

[16] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. https://doi.org/10.1109/78.650093

[17] Vaswani, A., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 1-15.

[18] Vig, J. (2019) A Multiscale Visualization of Attention in the Transformer Model. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Florence, July 2019, 37-42. https://doi.org/10.18653/v1/P19-3007

[19] Bahmani, A., *et al*. (2021) Hummingbird: Efficient Performance Prediction for Executing Genomic Applications in the Cloud. *Bioinformatics*, btab161.

[20] Nikolaou, C. and Almirantis, Y. (2005) "Word" Preference in the Genomic Text and Genome Evolution: Different Modes of n-Tuplet Usage in Coding and Noncoding Sequences. *Journal of Molecular Evolution*, **61**, 23-35. https://doi.org/10.1007/s00239-004-0209-2

[21] Huimin, X. and H. Bailin, (2002) Visualization of K-Tuple Distribution in Procaryote Complete Genomes and Their Randomized Counterparts. *Proceedings IEEE Computer Society Bioinformatics Conference*, Stanford, 14-16 August 2002, 31-42.

[22] Buenrostro, J.D., *et al*. (2013) Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nature Methods*, **10**, 1213-1218. https://doi.org/10.1038/nmeth.2688

[23] Bartlett, A., *et al*. (2017) Mapping Genome-Wide Transcription-Factor Binding Sites Using DAP-Seq. *Nature Protocols*, **12**, 1659-1672. https://doi.org/10.1038/nprot.2017.055

[24] Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A Census of Human RNA-Binding Proteins. *Nature Reviews Genetics*, **15**, 829-845. https://doi.org/10.1038/nrg3813