赵宇轩, 周阳, 丁学明. 金字塔型生成网络在蛋白质序列设计中的应用[J]. 智能计算机与应用,2024,14(12):124-132. DOI:10.20169/j.issn.2095-2163.241217

金字塔型生成网络在蛋白质序列设计中的应用

赵宇轩[1], 周 阳[2], 丁学明[1]
(1 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2 广西大学 计算机工程学院, 南宁 530007)

摘 要: GAN 网络在生成领域的研究与应用越来越成熟,但在模式崩溃问题和模式丧失问题上仍未得到很好的解决。本文提出了一个端到端的金字塔型多层 GAN 网络(Multi-Scale Generated Adversarial Network,MuSNET),以多层网络相链接的结构分担不同尺度的生成与优化任务,同时设计了全新的前馈特征提取模块,可以更好地捕获局部与全局之间的特征关系与相互作用,通过前馈方式加强输入信息,大幅提高了生成的多样性和模式稳定性。研究结果表明本模型在蛋白质序列设计领域中有着很好的表现。MuSNET 采用片段到整体的方式,以序列片段为基础,通过挖掘蛋白质序列进化耦合关系生成全新的完整序列。生成的序列符合天然序列同源性,并具有极高多样性,其分布特征、结构与功能位点也与天然序列一致,说明了 MuSNET 在保证模式稳定的同时,能有效捕获序列间特征关系,具有较高性能,对药物研发、靶点预测等领域有着重要意义。

关键词: 深度学习; 生成网络; 金字塔型; 蛋白质序列设计; 耦合关系
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2024)12-0124-09

Application of pyramidal generative network in protein sequence design

ZHAO Yuxuan[1], ZHOU Yang[2], DING Xueming[1]
(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 School of Computer Engineering, Guangxi University, Nanning 530007, China)

Abstract: Deep learning research and applications of GAN networks in the field of generation are becoming more mature. However, the problem of mode collapsing and mode dropping has not been well resolved. This paper develops an end-to-end pyramid-style multi-layer GAN network (Multi-Scale Generated Adversarial Network, MuSNET), which utilizes a multi-layer network architecture to handle generation and optimization tasks at different scales. A novel feed-forward feature extraction module is designed to better capture the feature relationships and interactions between local and global regions. By enhancing the input information through feed-forward propagation, MuSNET significantly improves the diversity and mode stability of generated samples and demonstrates excellent performance in protein sequence design. MuSNET adopts a fragment-to-whole approach, where new complete sequences are generated based on the coupling relationships uncovered from protein sequence evolution. The generated sequences conform to natural sequence homology and exhibit high diversity. Their distribution characteristics, structure, and functional sites are consistent with natural sequences, indicating that MuSNET ensures mode stability while effectively capturing the feature relationships between sequences. MuSNET performs with high performance, which is of significant importance in fields such as drug discovery and target prediction.

Key words: deep learning; generative network; pyramidal network; protein sequence design; coupling relationship

0 引 言

对抗生成网络[1](Generative Adversarial Network,GAN)是近年来深度学习领域中备受关注的一种强大的生成模型,通过 2 个互相博弈的神经网络生成逼真样本。GAN 在图像合成、文本生成、视频处理等领域取得了显著的成果,已成为机器学习研究中的一个热门话题。然而,许多现有的 GAN 模型往往存在模式崩溃和生成样本多样性不足等问题,限制了实际应用中的效果和应用广度,并且在蛋

基金项目: 国家自然科学基金(11502145)。
作者简介: 赵宇轩(1999—),男,硕士研究生,主要研究方向:深度学习; 周 阳(2003—),男,本科生,主要研究方向:机器视觉。
通信作者: 丁学明(1971—),男,博士,副教授,主要研究方向:深度学习,系统辨识,智能控制,电机控制等。Email:xuemingding@usst.edu.cn。
收稿日期: 2023-07-12

白质序列设计领域的研究也仍有很大提升空间。蛋白质序列可以决定其三维结构、理化性质和分子功能[2]，蛋白质序列设计的目标是设计具有特定性质或模式的全新序列[3]。得益于低成本的测序技术，数据库中存在大量序列可以用于模型的训练。深度学习与 GAN 模型能够从大规模序列中提取特征与耦合信息，用于设计稳定且多样的蛋白质[4]。因此，进一步开发和改进基于深度学习的 GAN 模型，提高蛋白质序列设计的效果和应用广度，能够推动蛋白质工程和生物技术的发展与突破，有着重大研究意义。

本文提出了一个端到端的金字塔型多层生成对抗网络 MuSNET 来产生蛋白质序列，利用蛋白质片段还原完整序列的方法学习氨基酸之间的复杂关系，并通过随机突变生成更多全新序列。MuSNET 的多层结构模拟自然界蛋白质的逐步进化过程，分担不同尺度任务的同时，扩大随机突变的影响，更大幅度地探索序列空间，提高生成样本的多样性。其前馈特征结构则保证了家族的主要模式和结构域不受影响。MuSNET 应用聚类后的序列对齐信息可以更容易地提取到家族序列特征，并且解决训练数据的样本均衡问题。采用序列片段到完整序列的生成方式可以更好地捕获天然序列样本局部到全局的相互作用与耦合关系，有效稳定了模式，弥补了当前研究的空白。

1　相关工作

目前 2 种主流蛋白质设计方法是自回归模型[5]（Auto Regression，AR）和 GAN 模型。自回归模型的代表是 Transformer 模型及其多种变体。Ferruz 等学者[6]用字母表示氨基酸，用字母组合表示氨基酸序列，提出 ProtGPT2 生成蛋白质序列，扩展当前结构数据库中的蛋白质。Moffat 等学者[7]采用 Transformer Decoder 架构，在一组人工合成序列上训练生成模型，并成功生成了具有有序结构的序列。Hesslow 等学者[8]提出了有 12 亿参数的 RITA 超大模型，并且证明了自回归模型随着规模的增大，模型效果及表现都有着明显提升。但是此类方法每次只能推理出一个标记（token），所以需要循环多次才能生成一串完整序列，在批量生成及扩展空间领域有着较大计算消耗，花费时间也较长。GAN 可以直接生成一条完整的序列，并且很早已被证明可以用于生成蛋白质序列的距离矩阵。Anand 等学者[9]在此基础上训练 GAN 模型，采用卷积、池化和上采样层来生成新的距离矩阵并以此重建序列，但是由于模型过于敏感、过程复杂、步骤长，容易产生错误。Repecka 等学者[10]开发了一个带有卷积层和注意力层的 ProteinGAN 来生成全新的、高度多样化的序列，并且在苹果酸脱氢酶家族（Malate dehydrogenase，MDH）上进行了实验，生成的序列基本具有 MDH 家族的主要结构域，保持了较低的相似性，成功扩展了 MDH 家族的序列空间。

2　数据来源与处理

为了构建基于家族的蛋白质生成模型，本文从 Pfam[11]数据库，以家族为单位构建数据集，同时对 3 个蛋白质家族进行了评估实验，分别是包含 17，271 条序列的白细胞介素-8 蛋白[12]（Interleukin-8，IL-8，PF00048），包含 8 454 条序列的肿瘤坏死因子蛋白[13]（Tumor Necrosis Factor，TNF，PF00229）和包含 18 097 条序列的小鼠双微体基因-2[14]（Mouse Double Minute 2，MDM2，PF02201）。

针对下载的家族数据集，进行如下的预筛选处理：先依据 4σ 原则，剔除长度低于 $\mu-4\sigma$（均值减 4 倍标准差）的序列，这些短序列很难包含完整的结构域和氨基酸间进化关系，保留下来可能会将模型引入错误方向。再利用 MMseqs2[15]，将一致性高于 50%的序列聚类到一个类中，每类取一条代表序列并通过插入 gap 的方式使其对齐到一致长度 $L_{\max}$，以此作为训练数据集。IL-8、TNF 和 MDM2 分别得到 381、376、340 条代表序列。聚类的方式可以去除高度重复的序列，使得各类数据之间的占比达到均衡，对解决模式崩溃问题有着关键作用。以代表序列作为数据来源，从每条代表序列中提取 5 个长度为 $L_{\max}/3$ 的片段构建数据集。这 5 个片段之间可能存在重叠，但将其加起来可以覆盖整条序列。在训练时，每个片段都是独立的训练样本。IL-8、TNF 和 MDM2 分别获得了 1 905、1 880、1 700 个训练样本。

3　模型设计

通常 GAN[16]网络由一个生成器和一个判别器组成，生成器目的是为了生成逼真的样本数据，判别器目的是为了区分真假数据。生成器接收概率分布 Z 后，生成虚假样本 x_g，将 x_g 和真实样本 x_r 输入判别器中，由判别器对二者进行区分。两者相互对抗，使得生成器不断生成更逼真的数据，判别器不断优化准确分类真假数据。通过这个对抗过程，GAN

不断地从数据中学习到特征,从而生成更逼真的样本。

MuSNET 借用了 SinGAN[17] 网络的思想,由 4 层逐级扩大的 GAN 网络堆叠组成,每层 GAN 网络具有相同的模型结构,但其中的维度参数会逐级扩大,如图 1 所示。4 个生成器依次链接,前一生成器的输出将作为后一生成器的输入,以实现序列的逐步扩张。每个生成器由 4 个残差空洞卷积特征提取模块(Residual Atrous Convolutional Feature Extractor Block,简称 RACF)和 1 个前馈推理模块组成,每个判别器由 4 个残差空洞卷积特征模块和 1 个二维卷积层组成。

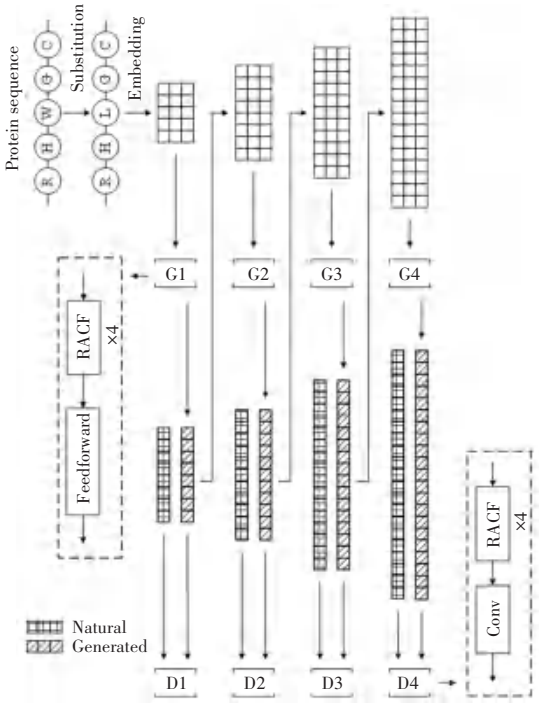


图 1 MuSNET 网络结构图

Fig. 1 Network architecture of MuSNET

3.1 序列编码嵌入

蛋白质序列由氨基酸排列组成,每个氨基酸对应一个字母代码,但此类文本序列无法被模型处理和识别,因此需要将其转换成模型可以处理的向量形式。MuSNET 使用独热编码(One-hot Encoding)方法实现序列编码嵌入[18],将每个氨基酸编码成长度 21 的一维向量,21 表示 20 种常见氨基酸和 1 个用于对齐的无意义标识符(gap)。在这个向量中,只有氨基酸对应位置的值为 1,其余位置都为 0,由此将长度为 L 的蛋白质序列编码成 $21\times L$ 的特征矩阵。One-hot 是一种计算机领域常用的编码方式,在蛋白质序列分析领域有着很好的表现,在实现数值化和特征化表示的同时,避免了氨基酸间的大小关系问题,更方便模型进行特征提取和分析,如图 2 所示。

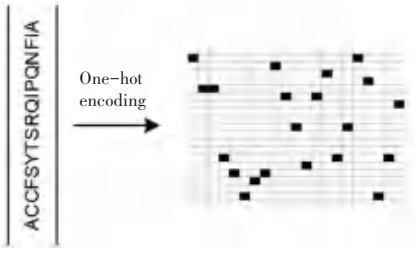


图 2 独热编码过程

Fig. 2 One-hot encoding process

3.2 RACF 模块

RACF 模块如图 3(a)所示,通过普通卷积、空洞率为 3 的空洞卷积[19]和空洞率为 5 的空洞卷积三种不同方式,分别提取多个通道特征,并通过平均池化层进行特征融合。空洞卷积可以扩大感受野,以更好地挖掘远距离氨基酸之间的相互关系,采用 2 种不同空洞率可以捕捉到不同尺度的序列特征,更全面地表示蛋白质序列的信息,提高模型对不同尺度特征的感知能力。

RACF 模块还借用了 ResNet[20] 的残差前馈思想,将卷积处理前的原始信息通过一条前馈通道叠加到池化特征上,以加强原始序列的编码特征,保留天然序列的基本架构,在生成序列过程中还可以强化单点突变造成的影响力。

RACF 模块中 3 个尺度的特征提取表达式如下:

$$\begin{cases} F_1 = Conv_{3,3,0}(x_{\text{input}}) \\ F_2 = Conv_{3,3,3}(x_{\text{input}}) \\ F_3 = Conv_{3,3,5}(x_{\text{input}}) \\ x_{\text{output}} = LR(BN(AVGP(F_1 + F_2 + F_3) + x_{\text{input}})) \end{cases} \tag{1}$$

其中,x_{input} 和 x_{output} 分别表示模块的输入特征和输出特征;$Conv_{i,j,d}$ 表示卷积层操作,这里 i,j 表示卷积核的大小,d 表示空洞率;$AVGP(\cdot)$ 表示平均池化层操作;$BN(\cdot)$ 表示批归一化操作;$LR(\cdot)$ 表示 LeakyReLU 激活函数。

生成器的前馈推理模块由卷积层、全链接层和 Softmax 层组成,如图 3(b)所示。先通过卷积将高维特征降至一维后,通过全连接层实现由短到长的序列生成。全连接层也称线性层,将前一层的全部神经元与后一层全部神经元相连并加权求和,在此依据卷积得到的特征矩阵推断新序列中每个氨基酸

的可能性大小。Softmax 将推理输出归一化，使得每列元素和为 1，列中的每个元素即为对应氨基酸的概率值。其中，每列最大值的位置对应的氨基酸即为该列生成的氨基酸，与序列的 One-hot 编码矩阵相对应。前馈推理模块表达式如下所示：

$$\begin{cases} x_{\text{output}} = Softmax(FC_{L1,L2}(Conv_{3,3,0}(x_{\text{input}}))) \\ x_{\text{output}} = Softmax(FC_{L1,L2}(Conv_{3,3,0}(x_{\text{input}}))) \\ FC_{L1,L2}(x_{21\times L1}) = x_{21\times L1} \cdot \omega_{L1\times L2} + \beta_{1\times L2} \end{cases} \tag{2}$$

其中，$FC_{L1,L2}(\cdot)$ 表示全连接层，$L1$，$L2$ 分别表示输入序列的长度和输出序列长度；$\omega_{L1\times L2}$ 和 $\beta_{1\times L2}$ 分别表示全链接层中每个氨基酸之间的权重和偏置；$Softmax(\cdot)$ 表示归一化。

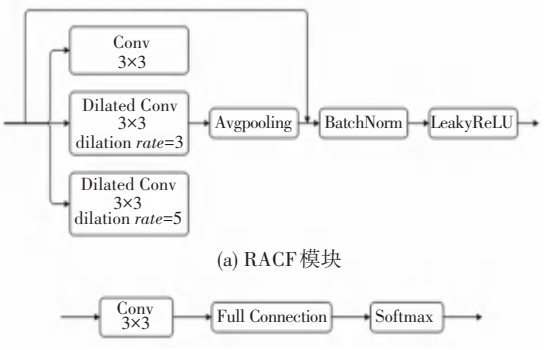


图 3　残差空洞卷积特征提取模块和前馈推理模块

Fig. 3　Residual atrous convolutional feature extractor block and feedforward block

3.3　生成器与判别器

MuSNET 有 4 层 GAN 网络，每层有一个对应的生成器，每个生成器都由 4 个 RACF 模块和 1 个前馈推理模块串联组成，如图 4(a)所示。生成器中的 4 个 RACF 模块有着相同的结构，但其中用于主要提取特征的卷积层的维度逐步增加，用于更好地让模型理解逐渐抽象化的特征数据。每个生成器有着相同的结构，但前馈推理模块中全连接层的参数受输入输出序列长度的影响而存在不同，越靠后的生成器参数量越大，以避免欠拟合的发生。

MuSNET 中的 4 个判别器有着完全一致的结构，由 4 个 RACF 模块和 1 个卷积层组成，如图 4(b)所示。与生成器原理一致，判别器中的 4 个 RACF 模块同样逐步增加卷积层的维度，而最后的卷积层将高维通道特征整合至单通道特征矩阵中。MuSNET 参考了 WGAN-GP 的方式，判别器直接输出矩阵作为序列在特征空间的分布，并判断真伪序列差距，详细内容见 5.1 节。

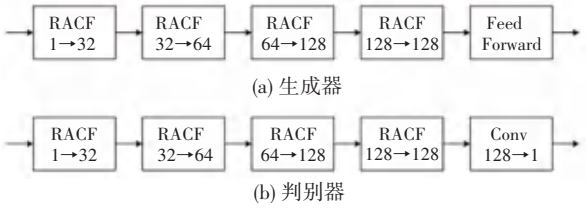


图 4　生成器与判别器结构图

Fig. 4　Structure of generator and discriminator

4　模型训练与生成

MuSNET 采用将天然序列片段恢复至完整序列的方式[21]，来学习序列中氨基酸之间的相互关系。MuSNET 的 4 层 GAN 网络采用逐层训练的方式，在训练过程中，先训练网络的第一层，然后固定第一层的参数，将第一层生成器 G_1 的输出作为第二层生成器 G_2 的输入，再训练第二层，以此类推，直到训练完所有层。在每个训练阶段，只有当前层 GAN_i 的参数是可调整的，而前面层 $GAN_{1\cdots i-1}$ 的参数被固定为之前训练好的值，后面的层 $GAN_{i+1\cdots 4}$ 则完全不参与训练。这种训练方式旨在逐步引入网络的复杂性，以更好地优化每一层的参数。通过逐层训练，每一层可以逐渐适应前一层的输出，从而提高整个网络的性能，还可以减少计算负担。

4.1　损失函数与超参设置

在本文的任务中，生成序列和原序列相差太大必然导致同源性的降低，相差太小又会导致缺乏多样性的缺失，因此在网络训练时需要平衡好同源性和多样性这两个方面的要求。GAN 网络优化方式参考了 WGAN-GP[22] 的方式，将判别器视作一个“打分模块”，但输出分数是一个矩阵形式的数据分布，再针对两分布间的差距评价生成结果的好坏。

生成器的损失函数由重构误差和对抗误差两部分组成。其中，重构误差旨在鼓励生成器生成与真实样本尽可能接近的样本，负责解决同源性问题；对抗误差旨在鼓励生成器生成逼真的样本使其能够欺骗判别器，负责解决多样性问题。重构误差是生成器 G_i 输出的概率矩阵和真实序列编码矩阵之间的均方误差(Mean Squared Error Loss，*MSELoss*)，对抗误差是判别器对生成样本打分的均值取相反数，将 2 种误差加权求和得到生成器的误差，并反向传播优化网络参数。*MSELoss* 是将矩阵各位置差值的平方和取平均，关注生成序列的不变性；对抗误差定义为生成样本的平均得分的相反数，是为了最大化生成样本的得分，使生成器能够更好地欺骗判别器，从而提高生成样本的质量。通过调整两误差之间的

权重,可以同步调整同源性和多样性之间平衡关系。生成器损失函数如下所示:

$$\begin{cases} L_G = \sigma_{\text{recon}} \cdot L_{\text{recon}} + \sigma_{\text{adv}} \cdot L_{\text{adv}} \\ L_{\text{recon}} = MSELoss(X_g, X_r) = \dfrac{1}{n}\sum (X_g - X_r)^2 \\ L_{\text{adv}} = -E[D(X_g)] \end{cases} \tag{3}$$

其中, X_g 表示生成样本; X_r 表示真实样本; σ_{recon} 和 σ_{adv} 分别表示两部分损失的权重,在此分别为 1 和 0.1; $E[\cdot]$ 表示计算均值; $D(\cdot)$ 表示判别器计算过程。

判别器的损失函数由 Wasserstein 距离和梯度惩罚(Gradient Penalty, GP)两部分组成。Wasserstein 距离用于衡量生成样本和真实样本分布之间的差异,GP 用于约束判别器的梯度,将 2 种误差加权求和得到判别器的误差,并反向传播优化网络参数。Wasserstein 距离是概率分布之间距离的度量,衡量将一种分布转换为另一种分布所需的最少工作量,而梯度惩罚用于约束判别器的梯度大小,使判别器满足 Lipschitz 连续性。判别器损失函数如下所示:

$$\begin{cases} L_D = \sigma_{\text{Wasserstein}} \cdot D_{\text{Wasserstein}} + \sigma_{GP} \cdot GP \\ D_{\text{Wasserstein}} = E[D(X_g)] - E[D(X_r)] \\ GP = E[(\mid \nabla_{\hat{x}} D(\hat{x}) \mid - 1)^2] \\ \hat{x} = \alpha \cdot X_r + (1 - \alpha) \cdot X_g \end{cases} \tag{4}$$

其中, $\sigma_{\text{Wasserstein}}$ 和 σ_{GP} 分别表示 2 部分损失的权重,在此分别为 1 和 0.1;“∇”表示梯度计算操作; $\hat{x}$ 表示生成样本和天然样本之间的差值; α 表示符合均匀分布的随机采样。

模型采用适应性矩估计优化器[23](Adaptive moment estimation, Adam),初始学习率设置为 1×10^{-5}, 每隔 500 个 *epoch* 衰减 1×10^{-1}。为节省训练时间并避免过拟合,训练结束条件采用 Early stopping 方式:将生成样本转换成对应的蛋白质序列,并每隔 100 个 *epoch* 计算全部生成序列和天然序列之间的汉明距离(Hamming distance),当 95% 以上的汉明距离达到 $0.05 \times L_{\max}$ 以下时停止训练。Early stopping 的条件如下所示:

$$\begin{cases} S_g[j] = argmax(x_g[:, j]) \\ D_{\text{Hamming}}(x_g, x_r) = \sum_{i=0}^{L_{\max}} (S_g[i] \neq S_r[i]) \end{cases} \tag{5}$$

其中, $argmax(\cdot)$ 表示寻找矩阵最大值对应的索引, S_g 和 S_r 分别表示生成序列和真实序列。

4.2 序列生成

新序列的生成采用单点突变方式[24],针对训练样本的序列片段,依次对每一个氨基酸进行 5 次随机的单点突变。将突变片段输入 MuSNET,经多层网络扩展后得到全新的完整序列。在自然界中,序列的氨基酸间存在相互依赖关系,又称共进化信息,当一个氨基酸发生突变时会导致其他氨基酸随之发生改变,而同一家族内的共进化信息存在一定的相似性。MuSNET 可以学习到家族内的氨基酸共进化信息,并将其作为生成序列的约束条件,面对某一氨基酸突变,MuSNET 会如自然界一般,根据共进化信息推动其他位点的氨基酸一起发生变化,以实现蛋白质稳定。

针对全部生成序列,进行数据过滤以更好地保证生成可靠性。将生成序列通过 ESMFold[25] 预测结构并计算 *pLDDT*0,并以 *pLDDT* 作为过滤条件。*pLDDT* 是在 0 ~ 1 范围内衡量每个残基的局部置信度,如果一条序列所有残基的平均 *pLDDT* 大于 0.5 即视为该序列较为可靠,并具有稳定结构。

针对 IL-8、TNF 和 MDM2 这 3 个家族,采用以上生成、过滤方式,分别得到了 67 791、62 567、89 681 条序列,并对比了生成序列和天然序列在同源性、序列潜在空间分布、残基分布、物理化学性质、结构、功能位点等多方面的相似性。

5 结果分析

蛋白质序列生成有 2 个基本评价指标、即同源性和多样性,MuSNET 可以从蛋白质家族序列比对中,挖掘家族同源性的基本规律,同时生成出与天然序列同源且更为多样的全新蛋白质序列,并且保留了进化特征、结构、功能等方面的特征,以下以 IL8 家族为代表展示结果分析。

5.1 同源性

同源性得分通过 HMMER[26] 工具进行计算和验证。HMMER 工具基于隐马尔可夫模型(Hidden Markov Model, HMM),对天然序列构建序列特征模型(Profile),并计算生成序列 X 在该模型下出现的概率 $P(X \mid profile)$,以此进行同源性分数(Homology Score, *HS*)的计算。一般当 $HS > 1$ 时,即可视为存在同源性;当 $HS > 10$ 时,视为具有极高同源性。得分计算方式如下所示:

$$HS = \log\left(\frac{P(X \mid profile)}{P_{\text{random}}(X)}\right) \tag{6}$$

其中, $P_{\text{random}}(\cdot)$ 表示随机序列在 profile 下的期

望概率，以此为基准来衡量生成序列在 profile 之间的匹配性。

将全部生成序列依次计算 *HS* 可以发现，MuSNET 展现出了很强的同源性能力，通过序列比对信息挖掘出了家族同源性规律，使生成的序列普遍具有极高同源性，也就意味着生成样本可以稳定达到天然样本的模式，具体统计见表 1。

表 1　序列同源性分析表

Table 1　Sequence homology analysis summary

家族	生成序列数	*HS* > 1	*HS* > 10	高同源/%
IL8	67 791	66 453	66 446	98.0
TNF	62 567	61 841	61 841	98.8
MDM2	89 681	88 135	88 135	98.3

5.2　多样性

多样性通过序列潜在空间分布来展现，通过将序列映射到潜在空间中，可以观察到天然序列和生成序列在空间中的分布情况，从而揭示序列之间的差异和相似性。先将天然序列和生成序列分别通过 MMseqs2 算法，按 50% 最小一致性为阈值进行聚类，天然序列得到 381 个小类，而生成序列得到 1 455 个小类，可见 MuSNET 生成的序列更为丰富多样，彼此间差距更大。通过一致性比对后，以 t-SNE[27]（t-Distributed Stochastic Neighbor Embedding）降维至二维潜在空间中。t-SNE 是一种常用的非线性降维技术，将高维数据转换成高斯分布，将低维数据转换成 t 分布，并采用最小化两分布之间 KL 散度的方式来进行降维。这种方式可以保留彼此之间的邻近关系，在低维空间中反映原始数据的局部结构信息与分布模式，其表达式如下所示：

$$\begin{cases} p_{ij} = \dfrac{\exp\left(-\dfrac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum\limits_{k\neq i}\exp\left(-\dfrac{\|x_i - x_k\|^2}{2\sigma_k^2}\right)} \\ q_{ij} = \dfrac{\exp(1+\|y_i - y_j\|^2)}{\sum\limits_{k\neq i}\exp(1+\|y_i - y_k\|^2)} \\ KL(P\parallel Q) = \sum\limits_{i\neq j} p_{ij}\log\dfrac{p_{ij}}{q_{ij}} \end{cases} \tag{7}$$

其中，x_i，x_j 和 x_k 分别表示数据点在高维空间的坐标。每个数据点都有多个维度。

经 t-SNE 降维后的结果如图 5 所示，MuSNET 生成的序列和天然序列的点在潜在空间中分布基本一致，两者相互重叠，可知生成序列在序列一致性比对方面基本与天然序列相似，可以被视作为同一家族。然而大部分天然序列都聚集到了少数大簇中，而生成序列可以存在许多较小的簇，相比之下具有更高多样性。可见 MuSNET 不是单纯地复制天然序列，更能够创造出具有独特变化的全新序列。此外，多样性的序列生成还有助于反映模型的鲁棒性和可靠性。由于生成的序列具有多样性，说明模型能够很好地适应不同的蛋白质序列变化和环境条件，提高了模型的泛化能力和适应性。生成序列簇无论大小，都分布于天然序列所包围的区域内，有效填补了天然序列间存在的空缺。并且生成序列的大簇基本靠近天然序列的大簇，可见两者在主体分布上的占比与数据流形也基本一致。通过生成多样的序列，能够探索更广泛的蛋白质序列空间，发现新的序列模式和功能特征，为蛋白质科学的发展提供了更多的创新思路。

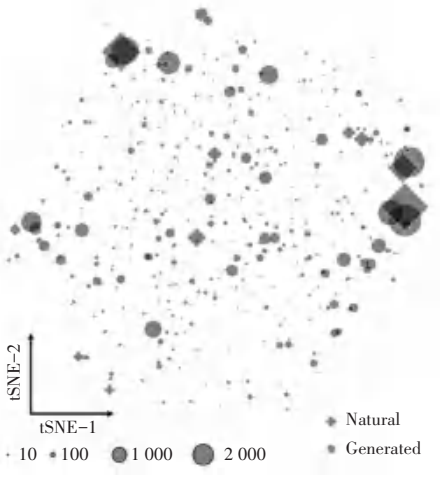


图 5　序列潜在空间分布

Fig. 5　Sequence latent space distribution

5.3　进化特征

MuSNET 的优越性体现在保证了同源性和多样性的同时，还保留了生成序列的进化特征，包括保守性、氨基酸对关联性、局部耦合性和全局耦合性[28]等。

保守性是指对应位置中，氨基酸发生改变的可能，若保守性强则说明该位置氨基酸基本不变或仅转变为同属性的相近氨基酸。生成的序列与天然序列的保守性在全局都保持了高度相似，保守性的高度一致说明了 MuSNET 能够准确地捕捉到了蛋白质序列的共性和保守区域，体现了模型在一阶序列进化特征上的特征提取能力，并且在生成序列时遵循蛋白质序列的生物学规律和限制，如图 6 所示。

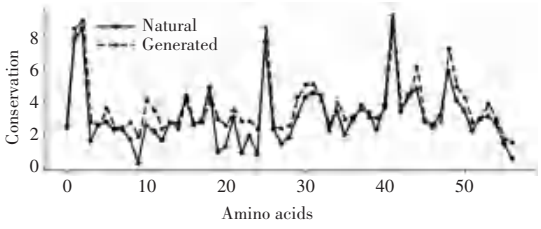


图 6 保守性对比图

Fig. 6 Conservation comparison

氨基酸对关联性通过不同种类氨基酸之间的距离分布,反映相互之间的关联性,以 Z_m 位置得分矩阵表示[10,29]。将已知序列与具有相同氨基酸频率的随机序列相比较,反映了氨基酸在已知序列和随机序列中的分布差异,从而表明 2 种氨基酸间的关联性。生成序列与天然序列的氨基酸对关联性有着高度相似性,如图 7 所示。MuSNET 学习到了不同氨基酸之间的相关联性,体现了 MuSNET 对不同氨基酸、不同元素编码之间的关系有着不错的表现。

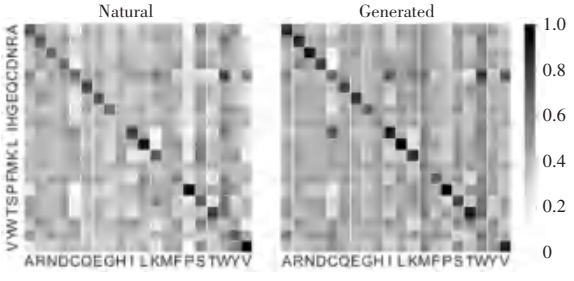


图 7 氨基酸对关联性对比图

Fig. 7 Amino acid pair association comparison

局部耦合性是指不同位置上的氨基酸之间存在的相互作用和依赖关系。这种耦合性的存在说明了蛋白质序列中的氨基酸不仅仅是独立存在的,而且对于蛋白质结构和功能的研究和应用具有重要的意义。生成序列与天然序列的局部耦合性有着高度相似性,如图 8(a)所示,说明 MuSNET 能够准确地捕捉到氨基酸之间的相互作用和依赖关系,体现了 MuSNET 在提取局部特征方面的优势。

全局耦合性是指不同位置上的氨基酸与整条序列之间存在的相互作用和依赖关系,反映了当一个氨基酸突变时,对整条序列造成的影响。生成序列与天然序列的全局耦合性有着高度相似性,如图 8(b)所示,可见 MuSNET 学习到了氨基酸与完整序列之间的耦合关系,体现了 MuSNET 在提取全局特征方面的良好表现。RACF 的引入使 MuSNET 能够同时提取到局部相邻和广域全局的特征。

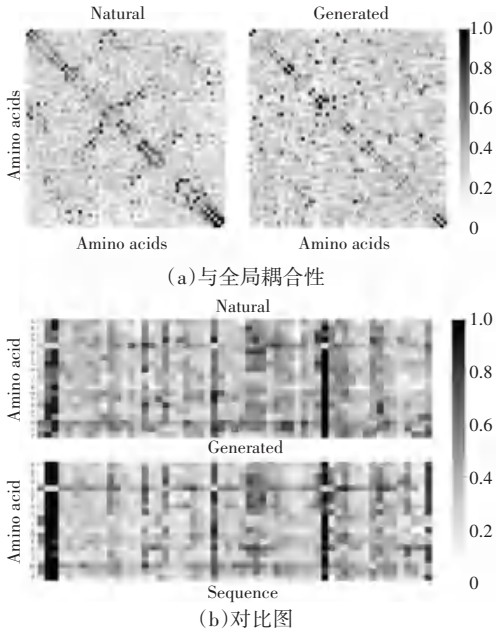


图 8 局部耦合性

Fig. 8 Local coupling

5.4 结构与关键位点

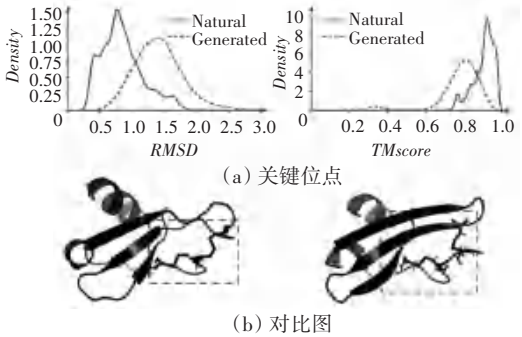
蛋白质家族通常由一个共同的祖先基因演化而来,因此都具有相似的氨基酸序列和结构域组成,以及类似的生物学功能。MuSNET 的生成序列能够捕获序列中蕴含的结构信息、关键位点与功能信息。本文比对了生成序列的预测结构与天然结构之间的 *TM - score* (Template Modeling score)和 *RMSD* (Root Mean Square Deviation)[30],分别从主干结构 *Cα* 原子和全原子两个角度比对结构相似度,*TM - score* 取值范围一般在 0~1 之间,一般高于 0.5 表示两结构具有较高相似性; *RMSD* 取值范围一般在 0~10 Å 之间,一般小于 2 Å 表示两蛋白质结构具有较高相似性,对比关系如图 9(a)所示。生成序列结构和天然序列结构有着极高相似度,且生成序列的结构完全由序列预测得来,可知生成序列中包含的结构特征信息与天然序列一致。

IL8 家族在结构上存在 4 个关键位点,即前文保守性中较高的 4 个半胱氨酸(Cystine,CYS),C2、C3、C26、C42。在天然序列中 C2 和 C26 之间、及 C3 和 C42 之间形成 2 对二硫键将蛋白质的前中后端链接,对稳定结构起着关键作用[31],在生成序列的结构中同样发现了这 2 对二硫键,如图 9(b)所示,左侧为天然序列结构,右侧为生成序列结构。生成序列中 2 对二硫键的位置和天然结构存在些许差别,但同样存在并起到稳定结构的作用。

IL8 家族是一种趋化因子,能够与 G 蛋白偶联

受体(G-Protein-Coupled Receptor,GPCR)结合并诱导白细胞趋向炎症部位,促进炎症反应和免疫细胞的活化和聚集,从而对机体的免疫防御和炎症反应产生影响[32]。通过 AF2 的 multimer 模式预测生成序列和 GPCR 蛋白的相互作用位置,发现生成序列可以通过将无规则卷曲(coil)部分伸入 GPCR 跨膜通道的方式进行结合,且结合方式、位置都与天然蛋白相似。

MuSNET 在训练过程中没有输入结构、相关位点或功能信息,但能够直接从序列中提取出相关表达并在生成具有一致特性的序列,体现了 MuSNET 出色的潜在信息挖掘能力。



(a) 关键位点

(b) 对比图

图 9　结构相似性

Fig. 9　Structure similarity

5.5　应用验证

本文致力于探索 MuSNET 模型及其生成的序列存在的应用研究价值,因此测试了其在 AlphaFold2[33](AF2)中的应用价值。AF2 是 DeepMind 开发的基于深度学习的蛋白质结构预测算法。算法针对输入序列先在天然数据库中搜索同源序列和结构模板,并以此作为模型的输入进行结构预测。当天然数据库中搜索到的同源序列或结构模板较少时,会导致预测结构出现重大偏差。

本研究将生成序列替代数据库中的搜索序列作为 AlphaFold2 的输入,且不再输入结构模板,仅依靠生成的同源序列进行结构预测。以生成序列作为输入得到的预测结构和天然结构有着极高的一致性,两者间的 $TM-score = 0.95$,$RMSD = 0.526$。由此可见 MuSNET 生成的序列可以在一定程度上取代天然序列,在蛋白质研究中发挥作用,证实了 MuSNET 在未来的后续研究中存在一定应用价值。

6　结束语

本文提出了一种端到端的金字塔型多层 GAN 模型 MuSNET,采用片段复原的方式进行训练,单点突变的方式进行生成,用于蛋白质序列生成。MuSNET 在长距离特征耦合问题、模式稳定问题、生成多样性问题等方面都展现出了优秀的性能。并且以 IL8 家族为例,从多个角度展示生成序列的研究成果,证明了 MuSNET 的多尺度特征提取能力与进化生成能力。该模型可以通过生成全新序列的方式,对天然蛋白数据库实现扩展,弥补天然蛋白在某些特性上的不足,有望对药物开发,生物催化剂性能改善等方面起到重要作用[34]。将深度学习和蛋白质序列设计相结合的思路,可以在丰富蛋白质数据库的同时,挖掘蛋白质的进化规律,推动人们对生命活动与演化的认识,值得进一步的探索和研究[35]。

参考文献

[1] STROKACH A, KIM P M. Deep generative modeling for protein design[J]. Current Opinion in Structural Biology, 2022, 72: 226-236.

[2] ANFINSEN C B. Principles that govern the folding of protein chains[J]. Science, 1973, 181(4096): 223-230.

[3] GE Qu, ZHU Tong, JIANG Yingying, et al. Protein engineering: From directed evolution to computational design [J]. Chinese Journal of Biotechnology, 2019, 35(10): 1843-1856.

[4] WANG Jingxue, CAO Huali, ZHANG J Z H, et al. Computational protein design with deep learning neural networks [J]. Scientific Reports, 2018, 8(1): 1-9.

[5] GREGOR K, DANIHELKA I, MNIH A, et al. Deep autoregressive networks[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR. org, 2014: 1242-1250.

[6] FERRUZ N, SCHMIDT S, HÖCKER B. ProtGPT2 is a deep unsupervised language model for protein design [J]. Nature Communications, 2022, 13(1): 4348.

[7] MOFFAT L, KANDATHIL S M, JONES D T. Design in the DARK: Learning deep generative models for de novo protein design[EB/OL]. (2023-02-01). https://www.biorxiv.org/content/10.1101/2022.01.27.478087v1

[8] HESSLOW D, ZANICHELLI N, NOTIN P, et al. Rita: A study on scaling up generative protein sequence models [J]. arXiv preprint arXiv,2205.05789, 2022.

[9] ANAND N, HUANG P. Generative modeling for protein structures [C]//Advances in Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2018, 31: 7494-7505.

[10] REPECKA D, JAUNISKIS V, KARPUS L, et al. Expanding functional protein sequence spaces using generative adversarial networks[J]. Nature Machine Intelligence, 2021, 3(4): 324-333.

[11] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: The protein families database in 2021 [J]. Nucleic Acids Research, 2021, 49(D1): D412-D419.

[12] HARTL D, LATZIN P, HORDIJK P, et al. Cleavage of CXCR1 on neutrophils disables bacterial killing in cystic fibrosis lung disease[J]. Nature Medicine, 2007, 13(12): 1423-1430.

[13] KERE J, SRIVASTAVA A K, MONTONEN O, et al. X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein[J]. Nature Genetics, 1996, 13(4): 409-416.
[14] KUSSIE P H, GORINA S, MARECHAL V, et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain[J]. Science, 1996, 274(5289): 948-953.
[15] STEINEGGER M, SöDING J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets[J]. Nature Biotechnology, 2017, 35(11): 1026-1028.
[16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.
[17] SHAHAM T R, DEKEL T, MICHAELI T. Singan: Learning a generative model from a single natural image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 4570-4580.
[18] OKADA S, OHZEKI M, TAGUCHI S. Efficient partition of integer optimization problems with one-hot encoding [J]. Scientific Reports, 2019, 9(1): 13036.
[19] ZHOU Zexun, HE Zhongshi, JIA Yuanyuan. AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images [J]. Neurocomputing, 2020, 402: 235-244.
[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.
[21] BERMAN D S, HOWSER C, MEHOKE T, et al. MutaGAN: A Seq2seq GAN framework to predict mutations of evolving protein populations[J]. arXiv preprint arXiv, 2008.11790, 2020.
[22] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs[J]. arXiv preprint arXiv, 1704.00028, 2017.
[23] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv, 1412.6980, 2014.
[24] BORNSCHEUER U T, POHL M. Improved biocatalysts by directed evolution and rational protein design[J]. Current Opinion in Chemical Biology, 2001, 5(2): 137-143.
[25] LIN Zeming, AKIN H, RAIVE R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model [J]. Science, 2023, 379(6637): 1123-1130.
[26] FINN R D, CLEMENTS J, EDDY S R. HMMER web server: Interactive sequence similarity searching [J]. Nucleic Acids Research, 2011, 39(suppl. 2): 29-37.
[27] MAATENV D L, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
[28] CHEUNG N J, PETER A T J, KORNMANN B. Leri: A web-server for identifying protein functional networks from evolutionary couplings [J]. Computational and Structural Biotechnology Journal, 2021, 19: 3556-3563.
[29] SANTONI D, FELICI G, VERGNI D. Natural vs. random protein sequences: Discovering combinatorics properties on amino acid words[J]. Journal of Theoretical Biology, 2016, 391: 13-20.
[30] ZHANG Yang, SKOLNICK J. TM-align: A protein structure alignment algorithm based on the TM-score[J]. Nucleic Acids Research, 2005, 33(7): 2302-2309.
[31] SU Yulin, YANG J C, LEE H, et al. The C-terminal disulfide bonds of Helicobacter pylori GroES are critical for IL-8 secretion via the TLR4-dependent pathway in gastric epithelial cells[J]. The Journal of Immunology, 2015, 194(8): 3997-4007.
[32] ARVANITAKIS L, GERAS-RAAKA E, VARMA A, et al. Human herpesvirus KSHV encodes a constitutively active G-protein-coupled receptor linked to cell proliferation[J]. Nature, 1997, 385(6614): 347-350.
[33] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.
[34] XIA Binbin, WANG Jun. Protein modeling and design based on deep learning[J]. Chinese Journal of Biotechnology, 2021, 37(11): 3863-3879.
[35] WU Qinglin, REN Yubin, ZHAI Xiaowei, et al. Protein sequence design using generative models[J]. Chinese Journal of Appied Chemistry, 2022, 39(1): 3-17.