arXiv:2304.03775v1 [stat.ML] 6 Apr 2023

具有灵活性保证的生物序列核

Alan N. Amin ALANAMIN@G.HARVARD.EDU
哈佛医学院

Eli N. Weinstein* EW2760@COLUMBIA.EDU
哥伦比亚大学

Debora S. Marks* DEBBIE@HMS.HARVARD.EDU
哈佛医学院 哈佛大学与麻省理工学院博德研究所

摘要

将机器学习应用于生物序列——DNA、RNA 和蛋白质——具有巨大潜力，有助于促进人类健康、环境可持续性和基础生物学理解。然而，许多现有的机器学习方法在这一问题定义域中效果不佳或可靠性不足。我们从理论角度研究这些挑战，并以核为切入点。基于核的方法无处不在：它们被用于预测分子表型、设计新型蛋白质、比较序列分布等。许多未明确使用核的方法仍会隐式依赖核，其中包括种类繁多的深度学习和基于物理学的技术。尽管针对其他类型数据的核在理论上已得到充分研究，但生物序列空间的结构（离散、长度可变的序列），以及生物学中的序列相似性概念，都带来了独特的数学挑战。我们对生物序列核在序列空间上逼近任意函数的能力，以及区分不同序列分布的能力进行了形式化分析。具体而言，我们确立了生物序列核成为通用核、具有特征性并使分布空间度量化的条件。我们表明，许多现有的生物序列基于核的机器学习方法并不满足这些条件，因此可能会严重失效。我们提出了简单且计算上易于处理的方法，以修改现有核，使其满足这些条件，从而为准确性和可靠性提供有力保证。我们的证明技术以具有离散质量的核理论为基础，并对其加以拓展。我们通过模拟和真实生物数据集展示了理论结果。

关键词： 核方法, 序列, 生物学, 非参数统计, 表示

1. 引言

考虑以下针对生物序列的机器学习问题。一位科学家想了解蛋白质序列与其荧光颜色之间的关系。他们合成了一个包含大量不同序列的库，并测量每种序列对应的颜色。他们的目标是根据序列预测颜色。一种可能采用的流行机器学习方法是高斯过程。高斯过程具有灵活性，能够表示不确定性，因此尤其适合帮助科学家设计具有特定颜色的新蛋白质。

*。贡献相同。

©2022年 Alan N. Amin、Eli N. Weinstein 和 Debora S. Marks。许可协议：
CC-BY 4.0，见 https://creativecommons.org/licenses/by/4.0/。

要在生物序列上构建高斯过程，必须指定一个定义在生物序列上的核 $k(X, Y)$ X, Y。人们已提出许多可能的选项，它们基于对序列相似性的不同理解。一种方法是逐位比较两个序列，并使用基于氨基酸生物物理特性的相似性矩阵（Schweikert等人，2008年；Toussaint等人，2010年）。另一种方法是比较每个序列中子串（k-mer）出现的频率（Leslie 等，2002年）。还有一种方法是根据两个序列之间所有可能的比对来评估序列相似性（Haussler，1999）。另一种方法则是根据深度神经网络生成的嵌入来比较序列（Yang 等人，2018年；Alley 等人，2019年）。所有这些方法都取得了一定的实证成效。

遗憾的是，正如我们将要展示的，对于所有这些不同的核，高斯过程模型都可能彻底失效：即使科学家将测量误差降至零，并扩大实验规模，在无限多条序列上收集无限量的数据，涵盖五彩缤纷的各种颜色，训练好的模型仍可能以很高的置信度预测每一条序列（无论其是否属于训练数据）都会产生荧光紫色蛋白质。根本问题在于，这些生物序列核都可能不是通用的：有些将序列与其属性联系起来的函数，它们根本无法描述。因此，基于这些核的方法可能并不可靠：它们在某些示例问题上表现良好，却会在新问题上失效，有时甚至一败涂地。

本文旨在为生物序列设计新的核，使其具有强有力的理论保证，避免出现不可靠的情况。具体而言，我们研究核的表达能力，也就是核能够或无法描述哪些函数（更专业地说，哪些函数属于该核关联的希尔伯特空间）。我们的目标是开发具有足够表达能力的生物序列核，从而确保使用这些核的机器学习方法能够在许多不同问题上可靠地发挥作用。

我们重点关注衡量核表达能力的三个关键指标：全域性、特征性，以及是否能使分布空间度量化。这里，我们简要介绍这些概念，并说明它们的实际意义。

在监督学习问题中，目标是从生物序列回归到另一变量，例如表型测量值。构建从基因型到表型的映射，对生物学、生物医学和生物工程的许多领域都很有价值，无论是推断古代微生物的食物来源、诊断遗传疾病，还是设计可持续催化剂等。监督学习的一种方法是使用线性回归等参数模型，但这类模型通常难以捕捉真实的序列到表型关系。更灵活的方法是采用基于核的回归方法，例如高斯过程或支持向量机。然而，核方法的灵活性未必能达到预期。为确保推断可靠，我们需要通用核，也就是说，它们能够表示序列上的任意函数。我们证明，许多现有的生物序列核并非通用核；同时，我们也提出了简单直接的修改方法，使它们成为通用核。使用我们提出的通用核的监督学习方法能够以可证明的方式描述任意基因型到表型映射，无论其复杂程度如何。

在无监督学习问题中，目标是学习序列上的一个分布。估计序列分布对生物学、生物医学和生物工程的许多领域都具有重要价值，例如重建古代免疫系统、预测未来的病毒突变、设计人源化抗体疗法等。在构建无监督

生成模型时，一个关键问题是模型的分布是否确实与数据的分布相匹配。比较序列分布的一种方法是使用汇总统计量，例如询问两个分布在序列长度、疏水性、预测结构等方面是否匹配。然而，即使这些分布在某些选定维度上相匹配，它们在其他方面也可能存在很大差异。一种更通用的比较方法是使用基于核的双样本检验或拟合优度检验，例如最大均值差异（MMD）（Gretton 等人，2012年）。为了确保推断可靠，我们希望使用具有特征性的核，这意味着当且仅当所比较的两个分布完全匹配时，MMD 才为零。我们证明，现有的核一般并不具有特征性；同时，我们还提供了可使其具有特征性的直接修改方法。使用这些修改后核之一的差异度量能够以理论上可证明的方式检测生物序列上两个分布之间的任何差异，无论这种差异多么细微。

基于核的方法不仅可用于评估生成模型，也可用于从一开始就学习生成模型。具体而言，可以通过最小化基于核的差异来训练无监督模型。这类方法尤其适用于似然无法解析求解的模型（也称为隐式模型），并已应用于生成对抗网络（GAN）、近似贝叶斯计算（ABC）等领域（Li 等人，2017年；Park 等人，2016年）。它们也适用于支持集未覆盖整个数据空间的模型，这一特性已被证明对实验设计和求积很有用（Chen 等人，2010年；Husz´ar 和 Duvenaud，2012年；Bach 等人，2012年；Pronzato，2021年）。这里的关键挑战在于，当我们通过优化最大均值差异损失来学习模型时，希望模型分布逐渐接近目标（数据）分布。即便核具有特征性，这一目标也未必能够实现。成功与否取决于该核能否度量分布空间。我们证明，现有的生物序列核通常不具备这一性质，并提出了具有这一性质的改进核。可以证明，使用我们提出的核最小化最大均值差异，将使模型分布与目标分布完全匹配。

在许多应用中，实践者会使用神经网络来替代核，或在核之外加以使用。例如，预测序列属性的一种常见方法基于半监督学习：首先，使用大规模无标签数据和深度神经网络学习从序列到表征的映射；然后，使用小规模有标签数据和高斯过程学习从表征到结果的映射（Yang 等人，2018年）。将从序列到表征的映射与高斯过程在表征上的核相结合，我们可以将这种方法理解为使用了一个作用于序列的核。我们在第10节分析这类核。更广泛地说，核与神经网络之间存在许多密切联系；核是从理论上研究深度学习方法的一种强大且常用的工具（Neal，1996年；Jacot 等人，2018年；Matthews 等人，2018年；Lee 等人，2017年；Simon 等人，2022年）。因此，我们的结果也为理论分析应用于生物序列的深度神经网络提供了一个起点。

在生物序列空间上研究核会带来独特的理论挑战。生物序列是由字符（核苷酸或氨基酸）组成的字符串，且通常在一个数据集中长度各异。因此，遵循此前关于生物序列的理论研究，我们将序列空间定义为字母表 $\mathcal{B}$ 上所有有限长度字符串 $\cup_{L=0}^{\infty}\mathcal{B}^L$ 的集合，其中例如，$\mathcal{B} = \{A, T, G, C\}$ 表示 DNA（Amin 等，2021年；Weinstein，2022年）。因此，生物序列空间是离散且无限的，而绝大多数以往关于核的理论和实证研究都在连续或有限空间上开展。

据我们所知，此前唯一研究无限离散空间中核灵活性保证的工作是 Jorgensen 和 Tian（2015年）的研究；他们只研究了少数几种对生物序列用途有限的核。

我们系统地研究了适用于生物序列的实用且广受欢迎的核的灵活性。我们的方法着重于确定某个核是否具有离散质量，即其关联的希尔伯特空间是否包含狄拉克函数（Jorgensen和Tian，2015年）。我们证明，这一性质足以保证该核是通用的、具有特征性的，并能度量化概率分布空间。由于欧几里得空间上的连续核不可能具有离散质量，我们的理论方法专用于离散空间。我们说明了核的多种组合与操作如何保留离散质量性质，并将这一理论用于提出对流行核的修改，使其具有离散质量。因此，我们为设计和构建具有强有力保证的生物序列核提供了一套强大的新工具。

本文结构如下。第2节介绍符号约定。第 3 节给出一个用于阐释本文结果及其实际应用方式的简单示例。第4节回顾相关工作。接下来的几节将建立用于证明核灵活性的关键理论工具。在第5节中，我们介绍所关注的几种核灵活性概念：全域性、特征性以及对分布空间的度量化。我们还引入具有离散质量的核这一概念，并证明它蕴含上述三种性质。第6节构建用于证明核具有离散质量的工具，尤其介绍能够保留离散质量性质的变换。在此基础上，我们研究了多种实践中使用的生物序列核；对于每一种核，我们都发现常见的核选择不具备所需性质，并提出了具备这些性质的替代方案。第7节考察逐位置比较序列的核，包括使用汉明距离的核。第8节考察比对核，这类核通过考虑序列之间所有可能的比对来比较序列。第9节考察 k-mer 谱核，这类核根据序列的子串（k-mer）集合来比较序列。第10节考察通过将序列嵌入欧几里得空间而构建的核。随后，我们在第11节中通过合成数据和真实数据展示研究结果，以及所提出核在实证表现上的提升。第12节作总结。

2. 符号约定

本节将说明本文后续部分所用的符号。

令 S 为一个具有离散拓扑的可数无限集合。在第 5 节和第 6 节中，我们将研究任意无限离散空间上的核所具有的保证，此时 S 可以是任意可数无限集合；在本文其余部分，我们将特别关注 S 为序列空间（定义如下）的情形。对于任意有限集合 B，我们将其基数记为 $|B|$。令 $\mathbb{N}$ 为自然数集合 $\{0, 1, 2, \dots\}$。我们还将 $\mathbb{1}(P)$ 定义为指示函数：若 P 为真，则其值为 1；若为假，则其值为 0。如果 $y, z \in \mathbb{R}$，则将它们的最大值记为 $y \vee z$，最小值记为 $y \wedge z$。

对于 $X \in S$，δ_X 指的是在 X 上取值为 1、在其他所有地方取值为 0 的函数或测度。我们将 $C_0(S)$ 称为 S 上在无穷远处消失的函数空间，并

$\|\cdot\|_\infty$ $C_0(S)$ 上的无穷范数。我们将 $C_C(S)$ 定义为 S 上仅在有限多个点非零的函数集合。我们将 $\mathcal{P}(S)$ 定义为 S 上的概率分布空间。

核是一个函数 $k\colon S\times S\to\mathbb{R}$，满足对于任意有限的 $Y_1,\dots,Y_N\in S$ 和 $\alpha_1,\dots,\alpha_N\in\mathbb{R}$，$\sum_{n,m}\alpha_n\alpha_m k(Y_n,Y_m)\geq 0$。我们在函数 $k_Y=k(Y,\cdot)$ 的线性组合上定义内积 $(\cdot|\cdot)_k$，使得 $(\sum_{n=1}^N\alpha_n k_{X_n}|\sum_{m=1}^M\beta_m k_{Y_m})_k=\sum_{n,m}\alpha_n\beta_m k(X_n,Y_m)$，其中 $X_n,Y_m\in S$ 和 $\alpha_n,\beta_m\in\mathbb{R}$。我们将相应的范数记为 $\|\cdot\|_k$。如果当 X_n 互不相同时，且 α_n 非零时，$\sum_{n=1}^N\alpha_n k_{X_n}\neq 0$，则称该核为严格正定核。我们将核 $\mathcal{H}_k$ 的再生核希尔伯特空间（RKHS）定义为在内积 $(\cdot|\cdot)_k$ 下的完备化所得的希尔伯特空间。我们可以将每个 $f\in\mathcal{H}_k$ 写成定义在 S 上的函数，具体为 $f(X)=(f|k_X)_k$。对于向量集合 $\{v_\lambda\}_\lambda\subset\mathcal{H}_k$，我们将 $\mathrm{span}\{v_\lambda\}_\lambda$ 定义为 $\{v_\lambda\}_\lambda$ 中元素的有限线性组合构成的集合。

设 μ 是 S 上的一个有符号测度，使得 $\int d|\mu|(X)\sqrt{k(X,X)}<\infty$。那么存在一个 $f\in\mathcal{H}_k$，使得对于所有 $g\in\mathcal{H}_k$，都有 $(f|g)_k=\int d\mu(X)g(X)$，并且 $\|f\|_k^2=\int d\mu(X)\int d\mu(Y)k(X,Y)$。这个 $\mathcal{H}_k$ 中的元素称为测度 μ 的“核嵌入”，记作 $\int d\mu(X)k_X$。

设 $\mathcal{B}$ 是一个有限集合，即“字母表”（例如，对于 DNA，它包含四种核苷酸；对于蛋白质，它包含二十种氨基酸）。我们将序列集合定义为 $\cup_{L=0}^\infty\mathcal{B}^L$，其中 $\mathcal{B}^0$ 定义为仅包含长度为零的序列 $\emptyset$ 的集合。对于序列 X 和数值 l，l'，我们将 $X_{(:l)}$ 定义为 X 中前 l 个字母，$X_{(l:l')}$ 定义为前 l 个字母之后的 $l'-l$ 个字母，而 $X_{(l:)}$ 定义为前 l 个字母之后的所有字母；它可能是空序列 $\emptyset$。我们称 $X_{(l)}$ 为 X 的第 l 个字母，计数从 0 开始。序列 X，Y 的拼接记作 $X+Y$。对于任意数值 l，将序列 X 与自身拼接 l 次，记作 $l\times X$。我们称 $d_H(X,Y)$ 为序列 X 和 Y 之间的汉明距离，也就是在两个序列末尾补上无限个终止符号 \$ 后，它们之间不匹配的位置总数。

3. 说明性示例

在全面阐述我们的理论结果之前，我们先通过一个示例说明其实际用途。我们考虑序列回归问题，其目标是预测序列的表型，例如荧光颜色、酶活性、结合强度等。我们将展示，使用经典生物序列核——滞后为 L 的汉明核（它是加权度核的一种）进行核回归时，可能完全无法拟合真实的序列到表型映射。随后，我们将介绍对该核的一项细微却不显而易见的修改；我们的理论保证，这一修改能够使核回归成功。

汉明核基于滑动窗口比较序列，统计窗口中的子序列（k-mer）完全匹配的次数（Schweikert等人，2008年）。更确切地说，我们考虑一种汉明核，它使用特征 $\Phi_{l,V}(X)=\mathbb{1}(X_{(l:l+L)}=V)$，用于指示位置 l 的 X 中的 L-mer 是否为 V。因此，该核为 $k_{\mathrm{H}}(X,Y)=(\Phi(X)|\Phi(Y))=\sum_{l=0}^\infty\sum_{V\in\mathcal{B}^L}\Phi_{l,V}(X)\Phi_{l,V}(Y)$，对于序列 X,Y，也就是说，它统计在 X 和 Y 中相同位置出现的 L-mer 的数量。请注意，当 $L=1$ 时，表征 $\Phi(X)$ 会简化为极为常用的 X 独热编码表示，此时该核的形式为 $k_{\mathrm{H}}(X,Y)=|X|\vee|Y|-d_H(X,Y)$，其中 d_H 是汉明距离。

然而，这种核可能无法描述序列与科学家所关注的表型之间的真实关系。原因在于，该核并非通用核，因此灵活性有限。为说明这一点，我们将证明， $\mathcal{H}_k$ 中的函数 f 只能取某些特定值。具体而言，函数 f 在序列集合 A_1 上取值，会受到它在子集 $A_2 \subset A_1$ 上取值的限制。只要数据集中序列的长度超过核中滑动窗口的长度 L，问题就会出现。定义 $A_1 = \{X\}_{|X|=L+1}$ 为所有长度为 $L+1$ 的序列构成的集合。定义 $A_2 = \{X + X_{(0)}\}_{|X|=L}$ 为长度为 $L+1$ 且末尾字母与开头字母相同（即 $X_{(0)}$）的序列集合。现在，在进行核回归时，我们拟合函数 $f = \sum_{n=1}^{N} \alpha_n k_{\mathrm{H}}(Y_n, \cdot)$。我们发现， f 必须满足以下约束：其在 A_1 上的平均值等于其在 A_2 上的平均值，

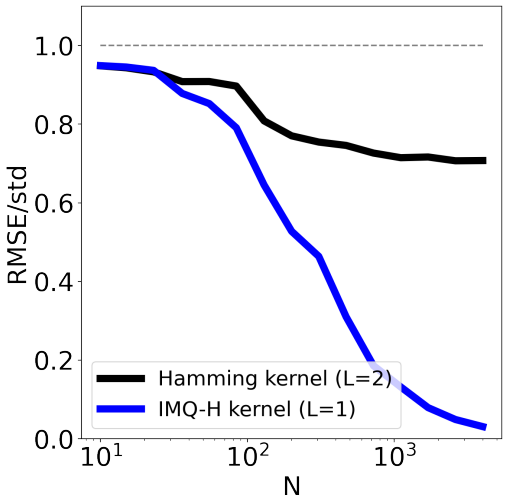
$$\begin{aligned}\frac{1}{|A_1|}\sum_{X\in A_1} f(X) &= \sum_{n=1}^{N}\alpha_n \left(\Phi(Y_n)\,\middle|\,\frac{1}{|A_1|}\sum_{X\in A_1}\Phi(X)\right)\\ &= \sum_{n=1}^{N}\alpha_n \left(\Phi(Y_n)\,\middle|\,\frac{1}{|A_2|}\sum_{X\in A_2}\Phi(X)\right) = \frac{1}{|A_2|}\sum_{X\in A_2} f(X).\end{aligned} \tag{1}$$

原因在于，对于每个 L-mer $V \in \mathcal{B}^L$ 和起始位置 $l \in \{0, 1\}$，在位置 l 具有 V 的序列所占比例，在 A_2 中与在 A_1中相同，即 $|\mathcal{B}|^{-L}$。因此，对于所有 l， V， $\frac{1}{|A_1|}\sum_{X\in A_1}\Phi(X)_{l,\,V} = |\mathcal{B}|^{-L} = \frac{1}{|A_2|}\sum_{X\in A_2}\Phi(X)_{l,\,V}$，从而得到公式 1。公式 1 的含义是，再生核希尔伯特空间（RKHS）中的函数受到约束，无法描述序列与其表型之间所有可能的关系；换言之，这表明该核不是通用核。请注意，这一论证也适用于更复杂版本的汉明核，例如考虑所有长度不超过 L的 K-mer 的核，或根据氨基酸的生物物理性质而非精确匹配来计算 k-mer 相似性得分的核（Schweikert等人，2008年；Toussaint等人，2010年）。

汉明核的有限灵活性可能导致严重的实际失败：模型并非只是在某些数据点上略有偏差，而是会对所有数据点都做出糟糕的预测。假设一位科学家测量了所有长度为 $L+1$ 的序列，得到数据集 $\{g(X)\colon\ X \in A_1\}$。他们发现，以相同字母开头和结尾的序列 $X \in A_2$ 的得分为 $g(X) = |\mathcal{B}| - 1$，而不满足这一条件的序列（即 $X \in A_1 \setminus A_2$）的得分为 $g(X) = -1$。我们假设不存在测量误差，即 $g(X)$ 就是表型的真实值。对于这个规模庞大且干净的数据集，使用汉明核进行核回归得到的最小二乘拟合结果竟然糟糕至极：$f(X) = 0$ 对所有 X 都如此。要看清这一点，请注意，最小二乘目标函数在以下位置取得最小值：$\alpha_1, \ \ldots, \ \alpha_N = 0$,

$$\begin{aligned}\frac{d}{d\alpha}\sum_{X\in A_1}\frac{1}{2}(\alpha k_{\mathrm{H}}(Y,X) - g(X))^2\Big|_{\alpha=0} &= -\sum_{X\in A_1} k_{\mathrm{H}}(Y,X)g(X)\\ &= \sum_{X\in A_1} k_{\mathrm{H}}(Y,X) - |\mathcal{B}|\sum_{X\in A_2} k_{\mathrm{H}}(Y,X)\\ &= |A_1|\left(\frac{1}{|A_1|}\sum_{X\in A_1} k_{\mathrm{H}}(Y,X) - \frac{1}{|A_2|}\sum_{X\in A_2} k_{\mathrm{H}}(Y,X)\right) = 0.\end{aligned}$$

图1：使用汉明核和 IMQ-H核进行最小二乘回归。 N 表示数据集的大小。性能通过均方根误差衡量，并按结果变量的标准差进行归一化。



因此，核灵活性不足是一个严重的实际问题，因为即使我们拥有大量干净的数据，也可能导致模型性能极差。

我们在第7节中的理论结果将给出一个简单直接的解决方案，即采用一种改进的核：逆多二次汉明（IMQ-H）核，

$$k_{\mathrm{IMQ-H}}(X,Y) = \frac{1}{(1+|X| \vee |Y| - (\Phi(X)|\Phi(Y)))^2}.$$

这种核的计算难度与原始汉明核一样低， $k_{\mathrm{H}}(X,\ Y)$ $=(\Phi(X)|\Phi(Y))$. 与原始核一样，它将具有相同位置上相同 k-mer 的序列视为彼此相似。因此，如果在某个具体问题中有生物学理由使用汉明核，那么同样的理由很可能也适用于 IMQ-H。[1]不过，我们将证明，I MQ-H 具有一项根本优势：它具有离散质量，这意味着它是通用的。若使用 IMQ-H 而非原始汉明核进行回归，我们就能保证可以描述任意序列到属性之间的关系。

我们通过模拟展示 IMQ-H核的优势。作为真实的序列到表型映射，我们考虑

$$g(X) = \max_b \sum_{l=0}^{\infty} \mathbb{1}(X_l = b).$$

1.例如，如果我们正在研究启动子，而其生物学活性取决于转录因子结合位点的排列方式，那么我们有充分的生物学理由相信，在相同位置具有相同K-mer的序列将具有相似的函数，因为转录因子通常会结合长度约为 $8-10$ 核苷酸的保守位点。在这种情况下，使用 $(\Phi(X)|\Phi(Y))$作为相似性得分是合理的，但生物学并没有特别的理由偏好 k_{H} 而不是 $k_{\mathrm{IMQ-H}}$。

它统计序列中最常见字母出现的次数。我们从长度为 4 的 DNA序列上的均匀分布中抽取数据点 X_1，X_2，.....，并记录每个数据点对应的无噪声观测值 $g(X_n)$。我们使用最小二乘核回归估计 g，并比较两种不同的核：参数为 $L = 2$ 的汉明核和参数为 $L = 1$ 的 IMQ-H核。直觉上，人们可能会认为 IMQ-H核的表现较差，因为它使用的滑动窗口更小。然而，实际上我们发现，随着数据不断增加，使用汉明核进行回归的误差会趋于一个较高值，而使用 IMQ-H核时，误差则收敛到零（图 1）。因此，我们的理论保证能够以几乎为零的计算成本带来显著的性能提升。

在接下来的章节中，我们将系统地扩展这个说明性示例中的思路。我们将介绍与广泛问题实际相关的理论保证，其适用范围远不止回归。我们将指出常用核（汉明核除外）存在的严重缺陷。此外，我们还将介绍替代核：它们保留了原有缺陷核所体现的相同生物学思想，计算上同样易于处理，却具备强有力的理论保证。

4. 相关工作

4.1 核的理论保证

关于核灵活性的理论研究成果十分丰富，其中许多工作聚焦于全域性、特征性以及使概率分布度量化等性质（Sripe-rumbudur 等人，2009年、2010年、2011年；Fukumizu 等人，2008年；Christmann 和 Steinwart，2010年）。这些结果奠定了重要基础，为各类核方法提供了有力保证；这些方法不仅包括回归方法，还包括双样本检验（Gretton 等人，2012年）、独立性检验（Gretton 等人，2007年）、样本质量评估（Gorham 和 Mackey，2017年）、隐式模型学习（Li 等人，2017年）、具有难以处理的归一化常数的模型学习（Dai 等人，2019年；Matsubara 等人，2022年）、因果推断（Singh 等人，2021年）、贝叶斯模型选择和数据选择（Weinstein 和 Miller，2023年）等。然而，几乎所有现有的灵活性保证都针对连续空间上的核。因此，将这些强大的核方法应用于一般的无限离散空间，尤其是生物序列时，仍缺乏理论依据。

对于实际应用中的核而言，情况尤其具有挑战性：它们必须捕捉真实世界的结构（即底层生物学结构），同时在计算上具有可处理性。事实上，对于实用的生物序列核，能否确立强有力的灵活性保证并非显而易见先验地：例如，在图的情形中，具有特征性的核至少与图同构问题一样难以计算，因此不存在多项式时间算法（Kriege 等人，2020年；G¨artner 等人，2003年）。换言之，可处理性与灵活性未必兼容。

Jorgensen和Tian（2015年）在无限离散空间上的核灵活性方面取得了开创性成果，探索了具有离散质量的核这一理念。我们大幅拓展了他们成果的适用范围和意义，具体而言：（a）将离散质量性质与机器学习应用中至关重要的灵活性概念联系起来；（b）研究与生物序列相关的核。更广泛地说，我们让Jorgensen

和Tian（2015年）的理论更“易于使用”：我们建立了更简便的途径来证明离散质量性质，并阐述了其实际意义。

4.2 生物序列核

尽管缺乏理论保证，关于生物序列核的实证研究文献依然丰富且历史悠久。我们将在后续章节详细介绍主要的生物序列核类别；此处先概述重要的整体趋势与挑战。首先，许多核（例如汉明核和加权度核）都按位置逐一比较序列，且往往依赖氨基酸相似性度量（对于蛋白质而言）（第7节）（Sonnenburg 等人，2007年；Toussaint等人，2010年）。这些逐位置比较核通常还依赖数据预处理，即进行多序列比对；这会要求数据集中的序列长度相同，并使不同位置之间的比较具有意义。这种预处理可能并不可靠，并会限制模型对未观测序列的泛化能力（Weinstein and Marks，2021年）。

一种不依赖比对预处理的方法是使用 k-mer 谱核，根据序列的 k-mer 含量对序列进行特征化（Leslie 等，2002）。另一种方法是使用比对核，它会明确考虑常见的生物突变（替换、插入和缺失），并对所有成对比对进行积分（Haussler，1999）。然而在实践中，比对核会表现出显著的“对角占优”现象：根据核参数的不同，序列要么都“非常接近”，要么都“相距甚远”，中间状态则不存在。解决这一问题的尝试大多未能成功；例如，有些已提出的方法并非非负定，或违反核方法的其他关键条件（Saigo 等人，2004年；Weston 等人，2003年）。我们将在第8节研究比对核和对角占优现象，并在第9节研究 k-mer 谱核。

近年来，利用深度神经网络进行表示学习的进展，嵌入核因此日益普及（Yang 等人，2018年）。在此，两个序列的核函数定义为：将标准欧几里得空间核（例如径向基函数核）应用于每个序列的表征。嵌入核可用于长度可变的序列，无需进行比对预处理，并且可以利用大规模无标签数据集来学习有意义的表征。然而，这些方法可能会将彼此毫不相关的序列嵌入到彼此接近的点上，尤其是当其中一个序列不在训练数据集内时。因此在实践中，研究人员通常将所考虑的序列限制在训练数据周围的一个邻域内，从而限制了该方法的适用范围（Yang 等人，2018年；Notin 等人，2021年；Stanton 等人，2022年；Detlefsen 等人，2022年）。我们在第10节研究嵌入核及其局限性。

5. 离散质量性质

本节将介绍离散质量性质，并说明它对核灵活性的影响。具体而言，我们将证明，如果一个核具有离散质量，那么它就具有以下性质：（1）通用（可用于描述序列空间上的任意函数）；（2）具有特征性（可用于区分任意序列分布）；（3）度量化分布空间（可用于优化一个序列分布，使其与另一个序列分布相匹配）。

通常，要为任何基于核的机器学习方法提供强有力的保证，都需要具备这三种性质中的一种或多种。后文中，我们将设计具有离散质量的生物序列核，因此这些核也满足上述三种性质。

如果希尔伯特空间包含数据空间 S 中所有点处的狄拉克函数，我们就称该核具有离散质量。请注意，在本节中，我们将 S 视为任意离散空间；稍后（从第7节开始），我们将专门讨论 S 为序列空间的情形。

定义 1（离散质量性质）。如果 k 满足 $C_C(S) \subset \mathcal{H}_k$，我们就称其具有离散质量，其中 $C_C(S)$ 是所有仅在有限多个点上非零的 S 上函数的集合。由于 $\mathcal{H}_k$ 是线性空间，这等价于对所有 $X \in S$，$\delta_X \in \mathcal{H}_k$。

在第5.1节中，我们将说明离散质量性质如何为函数近似提供保证；在第5.2节中，我们将说明它如何为区分分布提供保证。

5.1 通用核与函数逼近

使用核 k 进行回归时，人们用其再生核希尔伯特空间（RKHS）$\mathcal{H}_k$ 中的一个元素来拟合数据。RKHS 是如下形式的 S 上函数集合的闭包：$\sum_{n=1}^{N} \alpha_n k(\cdot, X_n)$，其中 $X_1, \ldots, X_N \in S$，且 $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$。回归的准确性在很大程度上取决于 $\mathcal{H}_k$ 中究竟包含哪些函数。粗略地说，$\mathcal{H}_k$ 越大，在大数据场景下回归方法就越准确、可靠。

如果再生核希尔伯特空间足够大，能够描述某个非常一般的集合中的任意函数，我们就称其具有通用性。具体而言，我们关注 C_0-全域性，即再生核希尔伯特空间能够逼近 C_0 中的任意函数；C_0 是所有在 S 上定义且在无穷远处消失的函数的集合（Sripe- rumbudur 等，2011）。

定义 2（C_0-全域性）：对所有 $X \in S$，设 $k(X, X) = 1$ 且 $k_X \in C_0(S)$。如果对于每个 $f \in C_0(S)$ 和 $\epsilon > 0$，都存在一个 $g \in \mathcal{H}_k$ 使得 $\|f - g\|_\infty < \epsilon$，则称 k 具有 C_0-全域性。

我们还将简要讨论 L^p-全域性，其定义方式相同，只是将$C_0(S)$替换为 $L^p(S, \mu)$，并将无穷范数替换为任意测度μ 下 S上的 p-范数。

常见的生物序列核是通用的吗？它们能描述任意基因型到表型的映射吗？许多都不能，因为它们不满足一个简单的标准：它们没有无限多个特征（Leslie 等，2002；Tsuda 等，2002；Jaakkola 等，2000）。

命题 3（在无限数据空间上，具有有限特征的核不是通用的）。考虑一个核 $k(X,Y) = (\phi(X) \mid \phi(Y))$，它由一个有限长度的特征向量 $\phi(X)$ 定义。如果 S 是无限的，那么该核不是 C_0-通用的（也不是 L^p-通用的）。

证明 在这种情况下，$\mathcal{H}_k$ 是有限维的，因为再生核希尔伯特空间 $\mathcal{H}_k$ 与 $\mathbb{R}^d$同构，其中d 是特征向量的长度。如果数据空间 S 是无限的，那么函数空间$C_0(S)$是无限维的；同样，对于任意 p，μ，函数空间 $L^p(\mu)$也是无限维的。因此，$\mathcal{H}_k$在 $C_0(S)$或 $L^p(\mu)$中都不可能稠密；$\mathcal{H}_k$ 中的每个函数都由其在 d个数据点上的取值决定。 ■

为作说明，我们应用这一命题来证明，一种广受欢迎的生物序列核——k-mer 谱核——并非通用。

例1（k-mer 谱核并非通用）。 *L-mer* 谱核将每种长度不超过 L 的子序列（*k-mer*）在序列中出现的次数作为特征（*Leslie* 等，*2002*）。其定义为 $k(X,Y) = (\Phi(X)|\Phi(Y)$，其中 $\Phi(X)$ 是一个向量，其分量 $\Phi(X)_V$ 表示 *k-mer* V 在 X 中出现的次数，对所有满足 $|V| \leq L$ 的 $V \in S$ 均如此。由于特征总数 $\sum_{l=1}^{L} |\mathcal{B}|^l$ 有限，因此该核不是通用核（命题 *3*）。

请注意，对于谱核的变体，同样的结论也成立：这类变体统计每种 k-mer V 在允许存在不匹配的情况下于 X 中出现的次数（Leslie 等，2004；Kuang 等，2004）。

尽管核要具有通用性，必须有无限多个特征，但这并非充分条件。在第 3 节中，我们证明了具有无限多个特征的汉明核并非通用核。相反，为了证明我们提出的新型生物序列核具有通用性，我们将证明它们具有离散质量。

命题 4 (具有离散质量的核是通用的)。 具有离散质量的核是 C_0-通用的（也具有 L^p-通用性）。

证明 $C_C(S)$ 在 $C_0(S)$ 中稠密；对于任意（可能是无限的）测度 μ，在 S 和 $\infty > p \geq 1$ 上的 $L^p(\mu)$ 中也稠密 ■

直观而言，具有离散质量的核的再生核希尔伯特空间包含每个序列对应的狄拉克函数；因此，根据线性性，将这些狄拉克函数相加，就能以任意精度逼近其他任何函数。

5.2 用于分布比较的特征核

除了用于回归，核也可用于比较分布。我们可以用最大均值差异（MMD）来衡量两个分布 μ 和 ν 之间的差异；它是核的再生核希尔伯特空间中某个函数的期望值之差的最大值（Gretton 等人，2012年）。为便于说明，本节还假设对于所有 $X \in S$，$k(X, X) = 1$。[2]

定义 5（最大均值差异）。 回顾测度 μ 的嵌入，记作 $\int d\mu(X) k_X$，它是 $\mathcal{H}_k$ 中的元素，满足 $(\int d\mu(X) k_X|g)_k = \int d\mu(X) g(X)$ 对所有 $g \in \mathcal{H}_k$ 都成立。最大均值差异（*MMD*）是两个分布 μ、ν 的嵌入之差的范数，

$$\mathrm{MMD}_k(\mu,\nu) = \max_{\{f\in\mathcal{H}_k:\|f\|_k\leq 1\}} E_\mu f - E_\nu f = \left\| \int d\mu(X) k_X - \int d\nu(Y) k_Y \right\|_k, \qquad (2)$$

其中 $\|f\|_k^2 = (f|f)_k$。

直观而言，再生核希尔伯特空间（RKHS） $\mathcal{H}_k$ 越大，就越能检测出 μ 与 ν 之间细微的差异。为了使最大均值差异（MMD）能够辨别任意 两个分布之间的差异，它必须具有特征性。

2. 这不会导致一般性的损失，因为在下文的论证中，我们可以将 $\mathcal{P}(S)$ 替换为 $\{\mu \in \mathcal{P}(S) \mid \int d\mu(X)\sqrt{k(X, X)} < \infty\}$。

定义 6 (特征核)。 我们称核 k 具有特征性，如果 $\mu \mapsto \int d\mu(X) k_X$在 $\mathcal{P}(S)$ 上是单射，其中该空间由 S 上的概率分布构成。

特征性意味着，$\mathrm{MMD}_k(\mu, \nu) = 0$ 当且仅当 $\mu = \nu$。例如，在实际应用中，特征性可确保基于核的双样本检验和条件独立性检验不会出现零检验力（Gretton 等人，2012年；Fukumizu 等人，2007年）。

许多常用的生物序列核并不具有特征性。例如，具有有限特征的核不仅不具备通用性，也不具有特征性。

命题 7 (具有有限特征的核在无限数据空间上不具有特征性)。考虑一个核 $k(X, Y) = (\phi(X) \mid \phi(Y))$，它由有限长度的特征向量 $\phi(X)$ 定义。如果 S 是无限的，该核就不具有特征性。

证明 证明见附录 A。 ■

与全域性一样，具有无限多个特征也不足以使核具有特征性。下面给出一个例子。

例 *2*（加权度核并非特征性核）。考虑第 *3* 节中的加权度核，以及集合 A_1 和 A_2。注意，该核具有无限多个特征。令 $\mu = \frac{1}{A_1} \sum_{X \in A_1} \delta_X$ 为 A_1上的均匀分布，并令 $\nu = \frac{1}{A_2} \sum_{X \in A_2} \delta_X$ 为 A_2上的均匀分布。于是 $\mu \neq \nu$，但 $\mathrm{MMD}_k(\mu, \nu) = 0$，因为分布嵌入满足 $\int d\mu(X) \Phi(X)_{l,V} = \frac{1}{|A_1|} \sum_{X \in A_1} \Phi(X)_{l,V} = |\mathcal{B}|^{-L} = \frac{1}{|A_2|} \sum_{X \in A_2} \Phi(X)_{l,V} = \int d\nu(X) \Phi(X)_{l,V}$，对所有 l, V 均成立。因此，MMD_k 并非区分分布的可靠方法。我们在附录 *B* 中通过模拟对此进行了说明。

最大均值差异不仅可用于比较两个给定的分布 μ 和 ν，还可作为优化目标，通过最小化该目标来寻找与 μ匹配的分布 ν。例如，我们可以通过寻找 $\mathrm{argmin}_\theta \mathrm{MMD}_k(q_\theta, \hat{q})$ 来训练模型 q_θ，其中 $\hat{q}$ 是训练数据的经验分布。这正是最大均值差异生成对抗网络（Li 等人，2017年）等方法背后的思路。最大均值差异也被用作多种其他领域的训练目标，包括实验设计、高维积分和近似贝叶斯推断方法（Pronzato 和 Zhigljavsky，2020年；Husz´ar 和 Duvenaud，2012年；Futami 等人，2019年）。

理想情况下，好的优化目标不仅应告诉我们是否已经得到正确答案，还应告诉我们是否正朝着正确的方向前进。因此，仅让最大均值差异告诉我们近似 ν 是否匹配目标分布 μ 还不够；随着我们通过寻找一个分布序列 $\nu_1, \nu_2, \ldots$使 $\mathrm{MMD}_k(\mu, \nu_n)$ 越来越小，我们还应得到对 μ越来越好的近似。要使这一点成立，核必须能够度量分布空间 $\mathcal{P}(S)$。

定义 8 (度量化 $\mathcal{P}(S)$). 我们称核 k 度量化 $\mathcal{P}(S)$，如果对于任意 $\mu \in \mathcal{P}(S)$和序列 $\nu_1, \nu_2, \ldots \in \mathcal{P}(S)$，均有最大均值差异$_k(\mu, \nu_n) \to 0$ 蕴含 $\nu_n \to \mu$。

请注意，对于连续空间上的核，类似性质有时称为“度量化弱收敛”（Sriperumbudur 等，2010）。然而，在 $\mathcal{P}(S)$ 上，弱拓扑等价于全变差拓扑（全变差收敛蕴含弱收敛；如果对所有 $X \in S$，$\nu_n(X) \to \mu(X)$，那么根据 Scheff´e 引理，$\sum_X |\nu_n(X) -$

$\mu(X)| \to 0$)。简而言之，能够度量化 $\mathcal{P}(S)$ 的核使我们能够可靠地将最大均值差异用作优化目标，而无法度量化 $\mathcal{P}(S)$ 的核则可能导致任意糟糕的解。

在优化生物序列时，表示学习方法在实践中颇受欢迎，因为与高维离散优化相比，低维连续优化更易处理（Yang 等人，2018年；Stanton 等人，2022年；Notin 等人，2021年）。然而，即使是能够保留大量序列空间信息、非常忠实的表示，也未必能够度量 $\mathcal{P}(S)$。因此，基于表示的嵌入核在优化序列分布时可能并不可靠。

示例 3（嵌入核并不保证能够度量 $\mathcal{P}$（S））。一位科学家关注的是生物序列上的分布 $p \in \mathcal{P}$（S），例如，生成模型预测会发出黄色荧光的蛋白质所构成的分布。他们希望从 中选择一个具有代表性的序列，在实验室中合成并测试。一种方法是选择 $\operatorname{argmin}_{X \in S} \mathrm{MMD}_k$（$p, \delta_X$），即按最大均值差异（*MMD*）衡量与 p 距离最近的序列。为了定义一个核，这位科学家使用一个嵌入 $F: S \to \mathbb{R}$，该嵌入是单射，因此每个序列都有独特的表示。

为了说明可能出现的问题，考虑一个简单情形：字母表 $\mathcal{B}$ 只包含一个字母 $\mathcal{B} = \{A\}$，序列空间为 $S = \{\emptyset, A, AA, \ldots\}$。我们将目标分布设为 $p = \delta_A$，即集中在长度为一的序列 A 上的点质量。为了定义核，我们考虑单射嵌入 $F(A) = 0$ 和 $F(n \times A) = 1/n$，其中 $n > 1$，并在嵌入空间中使用径向基函数核，使得完整的嵌入核为 $k(X, Y) = \exp(-(F(X) - F(Y))^2)$。直观上，由于径向基函数核在 $\mathbb{R}$ 上具有特征性，且嵌入 F 是一一对应的，因此嵌入核具有特征性（我们将在第*10*节中严格证明这一点）。

然而，这个嵌入核并不能度量 $\mathcal{P}(S)$。如果我们尝试相对于 X 最小化 $\mathrm{MMD}_k(\delta_A, \delta_X)$，会发现选择越来越长的序列会使目标值越来越接近零，$\mathrm{MMD}(\delta_A, \delta_{n \times A}) \to 0$，当 $n \to \infty$ 时。因此，尽管存在一个序列选择 $X = A$，使得近似 δ_X 与目标 $p = \delta_A$ 完全匹配，但优化最大均值差异并不会使我们得到正确答案。

为了设计出能够保证具有特征性并使 $\mathcal{P}(S)$ 成为度量的新的生物序列核，我们再次转向离散质量。

命题9（具有离散质量的核具有特征性，并使 $\mathcal{P}(S)$ 成为度量）。设 k是一个核，满足 $k(X, X) = 1$ 且 $k_X \in C_0(S)$ 对所有 $X \in S$。

$$k \text{ has discrete masses} \implies \mathrm{MMD}_k \text{ metrizes } \mathcal{P}(S) \implies k \text{ is characteristic.}$$

证明 第二个蕴含显然成立。现在设 k 具有离散质量，且$\mu, \nu_1, \nu_2, \cdots \in \mathcal{P}(S)$ 使得最大均值差异$_k(\mu, \nu_n) \to 0$。对于每个 $X \in S$, $\delta_X, -\delta_X \in \mathcal{H}_k$，因此根据方程2，我们有 $\|\delta_X\|_k^{-1} |\mu(X) - \nu_n(X)| \leq$最大均值差异$_k(\mu, \nu_n) \to 0$。因此，$\nu_n \to \mu$。 ■

因此，离散质量这一性质保证核具有通用性和特征性，并能度量化 $\mathcal{P}(S)$，使其成为适用于广泛应用的良好选择。[3] 我们

3. 请注意，离散质量性质比这三种性质都更强，因为存在一些核，它们是通用的并使 $\mathcal{P}(S)$ 成为度量，但不具有离散质量（附录C）。

将展示如何修改现有的生物序列核，使其具有离散质量，从而确保它们具有通用性和特征性，并能度量化 $\mathcal{P}(S)$。

5.3 退化示例

本文接下来将讨论如何设计具有离散质量的生物序列核。不过，在研究复杂的核之前，我们先考察两个简单但退化的具有离散质量的核，并说明它们为何不理想。一个具有离散质量的核是恒等核。

示例 4 (恒等核具有离散质量)。考虑如下形式的核$k(X,Y) = B(X)\delta_X(Y)$，对所有 X, $Y \in S$ 均成立，其中 $B:\ S \to (0,\infty)$ 是一个从序列空间映射到正实数的函数。这个核具有离散质量，因为对所有 X，均有 $B(X)^{-1}k_X = \delta_X$。问题在于，它会导致较差的泛化能力。例如，如果在高斯过程中使用这个核，那么对于训练集之外的点，其预测只取决于先验。

正如这个例子所示，核具有离散质量这一点本身，并不意味着它适合作为建模选择。相反，我们希望核不仅具有离散质量，还能体现生物学意义上的序列相似性。事实上，我们将遇到一些具有离散质量但对角占优的序列核，也就是说，它们的行为非常类似于恒等核。在这些情况下，我们会考虑在保留其离散质量的同时，对核进行修改，使其对角占优性降低。

构造具有离散质量的核的另一种方法，是假设序列空间为有限集，而非无限集。

命题 10 (有限空间上的严格正定核具有离散质量)。考虑有限的数据空间 S。如果核 k 是严格正定的，那么它具有离散质量。

证明 回顾一下，如果对于每个 $X \in S$，都存在某个 α，使得 $\delta_X(Y) = \sum_{X' \in S} \alpha_{X'} k(X', Y)$，那么该核具有离散质量。我们可以将 δ_X和 α视为长度为S的向量，从而得到方程 $\delta_X = K\alpha$，其中 K 是格拉姆矩阵，其元素对于所有 $X, Y \in S$均为 $K_{XY} = k(X,Y)$。核严格正定等价于其格拉姆矩阵严格正定，因此 K 可逆。于是，解 $\alpha = K^{-1}\delta_X$ 总是存在。

■

专注于有限的 S，从理论便利性的角度来看虽然颇具吸引力，但在实践中可能导致不可靠的方法。

示例 5 （长度有界的序列空间）。如果序列空间 S 仅包含长度小于某个最大值 $L_{\max}$ 的序列 X，即 $S = \cup_{L=0}^{L_{\max}} \mathcal{B}^L$，且核严格正定，那么该核具有离散质量。

问题在于，即使我们观察或优化的序列越来越长，也希望核方法始终可靠。例如，如果超过预先指定的长度尺度后出现新数据，我们不希望方法失效。即使我们选择 $L_{\max}$ 为天文数字般的大，使命题 *10* 在技术上得以满足，研究无限序列空间 $S = \cup_{L=0}^{\infty} \mathcal{B}^L$ 仍然有用，因为这会迫使我们考虑序列增长时方法的行为

的长度。换句话说，如果某种核方法在序列长度趋于无穷时的渐近行为很差，那么该方法在有限序列长度下也很可能表现不佳。

受这些问题的启发，我们旨在构造在无限序列空间上具有离散质量的核，并能够捕捉生物学意义上的序列相似性。

6. 描述和操作具有离散质量的核 s

我们已经看到，具有离散质量的核在逼近函数和区分分布方面具有有力保证。然而，迄今为止，只有少数与生物序列相关性有限的核被证明在无限数据空间上具有离散质量（Jorgensen和Tian，2015年）。本节将介绍可用于证明核具有离散质量的理论工具。第6.1节介绍与具有离散质量等价的条件；第6.2节介绍能够保留离散质量性质的核变换。稍后，我们将应用这些技术来设计具有离散质量的新型生物序列核。请注意，在本节中，S仍表示任意无限离散空间，不一定是序列空间。

6.1 离散质量性质的等价表述

本节将介绍离散质量性质的两种等价表述。

6.1.1 Conditions on the span of kernel embeddings

第一种表述对张成空间$\{k_Y\}_{Y\neq X}$施加条件，以保证 $\delta_X \in \mathcal{H}_k$。直观来说，考虑张成空间中的某个元素 $\sum_{Y'\in S}\alpha_{Y'}k_{Y'}\in$ 张成空间$\{k_Y\}_{Y\in S}$。如果 $\delta_X \in \mathcal{H}_k$，那么（$\delta_X|\sum_{Y'\in S}\alpha_{Y'}k_{Y'}\big)_k = \sum_{Y'\in S}\alpha_{Y'}\delta_X\ (Y') = \alpha_X$，也就是说，我们可以把 δ_X 看作一个函数，它将张成空间$\{k_Y\}_{Y\in S}$中的每个元素映射为 k_X前面的系数。要使这个函数存在，它必须满足：（1）定义良好；（2）有界。要使函数定义良好， k_X必须与 $\{k_Y\}_{Y\neq X}$线性无关。要使函数有界，还必须难以用张成空间$\{k_Y\}_{Y\neq X}$中的元素逼近 k_X 。这种直觉可以形式化如下。

命题 11。 设 $X \in S$，并将 $\mathcal{K}_X$称为张成空间$\{k_Y\}_{Y\neq X}$的闭包。 $\delta_X \in \mathcal{H}_k$ 当且仅当 $\mathcal{H}_k \neq \mathcal{K}_X$，换言之， $k_X \notin \mathcal{K}_X$。

证明 设 $\delta_X \in \mathcal{H}_k$。若 $Y \neq X$，则 $(\delta_X|k_Y)_k = \delta_X(Y) = 0$。另一方面，$(\delta_X|k_X)_k = 1$。因此，$k_X \notin \mathcal{K}_X$。

另一方面，设 $k_X \notin \mathcal{K}_X$。那么， $\mathcal{K}_X$ 的正交补恰好是一维的。令 $\phi : \mathcal{H}_k \to \mathbb{R}$ 为线性函数，将 $\mathcal{H}_k$ 投影到 $\mathcal{K}_X$的正交补上，并作缩放使得 $\phi(k_X) = 1$。那么，当 $Y \neq X$ 时，$\phi(k_Y) = 0$，因此，对所有 $Y \in S$， $\phi(k_Y) = \delta_X(Y)$。根据里斯表示定理，由于 ϕ 连续，存在一个 $f \in \mathcal{H}_k$，使得对所有 $Y \in S$， $f(Y) =(f|k_Y)_k = \phi(k_Y) = \delta_X(Y)$，因此 $f = \delta_X$ ■

这一结果的一个推论是，欧几里得空间上的连续核不可能具有离散质量。

命题 12 (欧几里得空间上的连续核不具有离散质量**)**。 如果核 k： $\mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ 是连续函数，则它不具有离散质量。

证明 考虑一个收敛于 $X \in \mathbb{R}^D$ 的序列 $Y_1, Y_2, \ldots \in \mathbb{R}^D$。则

$$\|k_{Y_n} - k_X\|_k^2 = k(X,X) + k(Y_n,Y_n) - 2k(X,Y_n) \to 2k(X,X) - 2k(X,X) = 0.$$

换言之， k_X 属于张成空间 $\{k_Y\}_{Y \neq X}$ 的闭包。 ■

因此，具有离散质量的核理论仅适用于无限离散空间，例如生物序列空间。

6.1.2 Conditions on the kernel's Gram matrix

离散质量的第二种表述源于 Jorgensen 和 Tian（2015 年）的开创性工作。我们在此简要重述他们结果的证明。其基本思想建立在命题 10 的基础上：该命题指出，在有限离散空间上，严格正定核具有可逆的格拉姆矩阵，因而具有离散质量。在无限空间上，我们考察定义在 S 的越来越大的有限子集上的一系列格拉姆矩阵的“可逆性”，以确保该核具有离散质量。

定理 13。 （***Jorgensen*** 和 ***Tian***，***2015*** 年）设 $X \in S$，且令 k 为 S 上的严格正定核。对于满足 $X \in B$ 的有限子集 $B \subset S$，定义由 B 索引的格拉姆矩阵 K_B，使得 $\underline{K_{B,\ Y,}}_{Y'} = k(Y,\ Y')$，其中 $Y,\ Y' \in B$。令 $C_{B,\ X} = \sqrt{\left(K_B^{-1}\right)_{X,\ X}}$。则，$\delta_X \in \mathcal{H}_k$ 当且仅当 上确界$_B C_{B,\ X} < \infty$，其中上确界取遍所有有限 $B \subset S$。在此情况下， $\|\delta_X\|_k =$上确界$_B C_{B,\ X}$。

证明 定义一个泛函 ϕ_X，其定义域为张成空间$\{k_Y\}_{Y \in S}$，并且满足 $\alpha_X k_X + \sum_{Y \neq X} \alpha_Y k_Y \mapsto \alpha_X$。注意，由于 k 是严格正定的，ϕ_X 定义良好。现在，若 $\delta_X \in \mathcal{H}_k$，则对所有属于张成空间$\{k_Y\}_{Y \in S}$的 $f \in$，有 $(\delta_X|f)_k = \phi_X(f)$，因此 ϕ_X有界。另一方面，若 ϕ_X有界，则可将其连续延拓到整个 $\mathcal{H}_k$ 上，并且根据里斯表示定理，它必定等于某个$f \in \mathcal{H}_k$对应的 $(\cdot|f)_k$。于是， $f(Y) = (k_Y|f)_k = \phi_X(k_Y) = \delta_X(Y)$，因此 $f = \delta_X$。我们将证明， C_B是 ϕ_X在张成空间$\{k_Y\}_{Y \in B}$上的限制的范数，结论便随之成立。

$\{k_Y\}_{Y \in B}$ 是一个有限维空间，以 $\{k_Y\}_{Y \in B}$ 为基，内积 $(v|w) = v^T K_B w$（对于 $v,\ w \in \mathbb{R}^B$，以此为基）。令 e_X 为 X 的指示向量，则 $\phi_X(v) = (e_X^T K_B^{-1}|v)$。最后，我们看到 ϕ_X 的平方范数为 $(e_X^T K_B^{-1}|e_X^T K_B^{-1}) = \left(K_B^{-1}\right)_{X,X} = C_B^2$。 ■

6.2 保持离散质量性质的变换

在本节中，我们将介绍如何基于现有的具有离散质量的核构造新的具有离散质量的核。这使我们能够构造出彼此相关的大型核族，并且这些核都具有灵活性保证。

6.2.1 Summing

首先，我们考虑具有离散质量的核的线性组合。核通常会被求和或积分（Duvenaud 等，2013）。例如，要构建一个回归模型

以便轻松学习多个长度尺度上的现象，可以将带宽参数取值不同的核相加，或对所有带宽进行积分。下面的结果表明，只要这个和中有一个核具有离散质量，那么整个求和后的核也具有离散质量。

命题 14. 设 Λ 是一个可测空间，且$(k_\lambda)_{\lambda\in\Lambda}$ 是一个核族。假设对于任意 $X, Y \in S$, $\lambda \mapsto k_\lambda(X,Y)$ 均可测。设 ν 是 Λ上的一个非零正测度，满足 $\int d\nu(\lambda) k_\lambda(X,X) < \infty$ 对所有 $X \in S$ 都成立，且 $\nu(\{\lambda \mid k_\lambda \text{ 具有离散质量}\}) > 0$，即 ν在具有离散质量的核上具有正质量。[4] 那么， $k_\nu = \int k_\lambda d\nu(\lambda)$ 是 S上的一个具有离散质量的核。

证明 我们考虑 k_ν 不具有离散质量的情形，并证明这会导致矛盾。根据命题 11，存在某个 $X \in S$ 和序列 $\mu_l = \delta_X - \sum_{Y\neq X} \alpha_{l,Y} k_Y$，其中 $\sum_{Y\neq X} \alpha_{l,Y} k_Y \in$张成$\{k_Y\}_{Y\neq X}$，使得 $\| \int k_{\nu,Z} d\mu_l(Z)\|_{k_\nu} \to 0$当 $l \to \infty$时成立。于是，

$$\begin{aligned}\left\| \int k_{\nu,Z} d\mu_l(Z) \right\|_{k_\nu}^2 &= \int d\mu_l(Z_1) \int d\mu_l(Z_2) k_\nu(Z_1, Z_2) \\ &= \int d\nu(\lambda) \int d\mu_l(Z_1) \int d\mu_l(Z_2) k_\lambda(Z_1, Z_2) \\ &= \int d\nu(\lambda) \left\| \int k_{\lambda,Z} d\mu_l(Z) \right\|_{k_\lambda}^2 .\end{aligned} \tag{3}$$

因此，根据 Fatou 引理，

$$\int d\nu(\lambda) \liminf_{l\to\infty} \left\| \int k_{\lambda,Z} d\mu_l(Z) \right\|_{k_\lambda}^2 = 0.$$

特别地，存在一个 λ，使得 k_λ 具有离散质量，且 $\liminf_{l\to\infty} \| \int k_{\lambda,Z} d\mu_l(Z)\|_{k_\lambda}^2 = 0$，这就产生了矛盾。 ■

一种特殊情况是对一组核求和。

推论 15. 设 $k_0, k_1, \ldots k_N$ 为核，且 k_0 具有离散质量，并且$\alpha_0, \ldots, \alpha_N > 0$。那么 $\sum_{n=0}^N \alpha_n k_n$ 是一个具有离散质量的核。

证明 设 $\nu = \sum_{n=0}^N \alpha_n \delta_n$。现在，根据上述命题， $\sum_{n=0}^N \alpha_n k_n = \int d\nu(n) k_n$ 具有离散质量。

■

6.2.2 Changing domains

接下来，我们考虑在序列空间的不同区域使用不同的核。以下结果表明，如果一个核在整个 S上具有离散质量，那么将其限制在 S的某个区域内时，也具有离散质量。此外，如果我们在 S的不同正交区域上分别有具有离散质量的核，就可以将它们组合起来，构造一个在整个 S上具有离散质量的新核。

4. 如果 $\{\lambda \mid k_\lambda$ 具有离散质量$\}$ 不可测，那么我们改为要求，对于所有 $C \supset \{\lambda \mid k_\lambda$， $\nu(C) > 0$ 具有离散质量$\}$。

命题 16。 设 $\{W_V\}_V$是 S 的两两不相交子集的集合，且满足 $S = \cup_V W_V$。如果 k 具有离散质量，那么将其限制在任意 W_V上时，也具有离散质量。另一方面，如果 k 在每个 W_V上受限时都具有离散质量，并且对于任意 $X \in W_V$、 $X' \in W_{V' \neq V}$，都有 $k(X, X') = 0$，那么 k 在 S上具有离散质量。

证明 首先假设 k 在 S上具有离散质量。核的一个标准性质（详见附录 D 中的命题 37）是：将核限制在定义域W_V上所得的希尔伯特空间，即 $\mathcal{H}_{k|_{W_V}}$，是原希尔伯特空间 $\mathcal{H}_k$中 $\text{span}\{k_Y\}_{Y \in W_V}$的闭包。现在考虑任意一个 $X \in W_V$。根据命题 11， k_X不在 $\text{span}\{k_Y\}_{Y \in S \mid Y \neq X}$的闭包中，因此它也不在 $\text{span}\{k_Y\}_{Y \in W_V \mid Y \neq X}$的闭包中。于是，再次应用命题 11 可得，$k$ 在 W_V上具有离散质量。

现在考虑这样的情况：将 k 限制在每个 W_V上时，它都具有离散质量。假设对于某个 $X \in W_V$， $\delta_X \notin \mathcal{H}_k$；我们将证明这会导致矛盾。根据命题 11，存在一个函数序列 $f_1 = \sum_{Y \neq X} \alpha_{1,Y} k_Y, f_2 = \sum_{Y \neq X} \alpha_{2,Y} k_Y, \cdots \in$，其张成空间为 $\{k_Y\}_{Y \neq X}$，并且当 $m \to \infty$ 时， $f_m \to k_X$。令 $f_{V,m}$ 为f_m 在 $\text{span}\{k_Y\}_{Y \in W_V}$上的正交投影，即 $f_{V,m} = \sum_{Y \neq X} \mathbb{1}(Y \in W_V)\alpha_{m,Y} k_Y$。注意，对于每个$m$，都有 $f_{V,m} \in \text{span}\{k_Y\}_{Y \in W_V \mid Y \neq X}$，因此 $f_{V,m}$ 属于将核限制在 W_V上所得的希尔伯特空间 $\mathcal{H}_{k|_{W_V}}$。定义 $f_{V,m}^{\perp} = f_m - f_{V,m} = \sum_{Y \neq X} \mathbb{1}(Y \notin W_V)\alpha_{m,Y} k_Y$。现在 $(f_{V,m}^{\perp} | f_{V,m})_k = 0$且 $(f_{V,m}^{\perp} | k_X)_k = 0$，因此

$$\|f_{V,m} - k_X\|_k^2 \leq \|f_{V,m}^{\perp}\|_k^2 + \|f_{V,m} - k_X\|_k^2 = \|f_m - k_X\|_k^2 \to 0.$$

换言之， k_X属于 $\text{span}\{k_Y\}_{Y \in W_V \mid Y \neq X}$ 的闭包。这意味着，限制在 W_V 上的核k 不具有离散质量，这与假设矛盾。 ■

6.2.3 Tensorizing

接下来，我们考虑核的张量化，使其能够应用于序列对。若我们有定义在 S上的两个核 k_1 和 k_2，则张量化后的核 $k_1 \otimes k_2$为 $k_1 \otimes k_2((X_1, X_2), (Y_1, Y_2)) = k_1(X_1, Y_1)k_2(X_2, Y_2)$，其中 $X_1, X_2, Y_1, Y_2 \in S$。例如，张量化核可用于确定两个随机变量是否相互独立（Fukumizu 等人，2007年）。张量化保留离散质量。

推论 17. 设 k_1和 k_2是定义在 S上且具有离散质量的核。则 $k_1 \otimes k_2$是定义在 S^2 上且具有离散质量的核。

证明 核的一个标准性质（附录 D 中的命题 36 详细说明了这一点）表明，如果 $f_1 \in \mathcal{H}_{k_1}$ 和 $f_2 \in \mathcal{H}_{k_2}$，则存在一个 $f_1 \otimes f_2 \in \mathcal{H}_{k_1 \otimes k_2}$，使得对于 $(X_1, X_2) \in S^2$，有 $f_1 \otimes f_2(X_1, X_2) = f_1(X_1) f_2(X_2)$。设 $X, Y \in S$。那么 $\delta_{(X,Y)} = \delta_X \otimes \delta_Y \in \mathcal{H}_{k_1 \otimes k_2}$。 ■

6.2.4 Tilting

接下来，我们考虑对核重新加权，以强调或弱化序列空间中的某些区域。更准确地说，我们考虑通过某个函数 A 对核 k 进行倾斜： $S \to (0, \infty)$，从而得到一个新核 $k^A(X, Y) = A(X)k(X, Y)A(Y)$，其中 $X, Y \in S$。对一个核进行倾斜的一个原因是

核 k 的做法是将其“归一化”，使得 $k^A(X,X) = 1$ 对所有 $X \in S$成立；这对应于倾斜函数 $A(X) = \sqrt{k(X,X)}^{-1}$。倾斜后仍保留离散质量性质。

命题 18. 如果 k 是 S 上的一个具有离散质量的核，那么 k^A 具有离散质量。

证明 核的一个标准性质（详见附录 D 中的命题 35）是：如果原再生核希尔伯特空间 $\mathcal{H}_k$ 中有一个函数 $f(\cdot)$，那么函数 $A(\cdot)f(\cdot)$ 属于倾斜核 $\mathcal{H}_{k^A}$对应的再生核希尔伯特空间，且范数相同，即 $\|f\|_k = \|Af\|_{k^A}$。换言之，$f \mapsto Af$ 是从 $\mathcal{H}_k$ 到 $\mathcal{H}_{k^A}$的等距同构。因此，对于任意 $X \in S$，我们有$A\delta_X \in \mathcal{H}_{k^A}$。我们总可以将再生核希尔伯特空间中的函数乘以一个有限标量。因此，$\delta_X = {}_A{}^1{}_{(X)} A(\cdot)\delta_X(\cdot) \in \mathcal{H}_{k^A}$。 ■

6.2.5 Reparameterizing alphabets

最后，我们来看一种针对生物序列的核的新型变换，它涉及对字母表 $\mathcal{B}$ 进行“重新参数化”。基本思路是将字母表 $\mathcal{B}$ 中的字母表示为独热编码向量，即将每个 $b \in \mathcal{B}$ 视为一个长度为 $|\mathcal{B}|$ 的向量，除一个位置外，其余位置均为零。因此，字母表 $\mathcal{B}$ 构成了 $\mathbb{R}^{\mathcal{B}}$ 的一组基。不过，我们也可以考虑 $\mathbb{R}^{\mathcal{B}}$ 的另一组基 $\tilde{\mathcal{B}}$。这组替代基会产生另一组序列 $\tilde{S} = \cup_{L=0}^{\infty}\tilde{\mathcal{B}}^L$。将 S 和 $\tilde{S}$ 视为向量集合后，$\tilde{}$ 就可以自然地将定义在 S 上的核扩展到 S 上。我们将看到，离散质量这一性质在基变换下保持不变：如果该核在 $\tilde{S}$ 上具有离散质量，那么它在 S 上也具有离散质量，反之亦然。（请注意，在本节及后续所有章节中，S 指的是序列集合，而不是任意的无限离散空间。）

更准确地说，我们必须定义如何将序列核应用于序列的向量编码。任意序列 Yin $S = \cup_{L=0}^{\infty}\mathcal{B}^L$ 都可以表示为独热编码，它由一个向量 $V \in \mathbb{R}^{|Y|\times\mathcal{B}}$ 组成，其中 $V_{(l)} \in \mathbb{R}^{\mathcal{B}}$ 是字母 $Y_{(l)}$的独热编码。如果向量 $V \in \mathbb{R}^{L\times\mathcal{B}}$ 是某个序列的独热编码，那么我们将其嵌入到 $\mathcal{H}_k$ 中定义为它所编码序列的嵌入。我们可以用以下公式表示：

$$k_V = \sum_{X\in S:|X|=L} \left(\prod_{l=0}^{L-1} V_{(l),X_{(l)}}\right) k_X. \tag{4}$$

因此，如果 V是 Y的独热编码，我们就能得到 $k_V = k_Y$。核并未改变；我们所做的只是改写它，以便嵌入向量编码的序列。

现在，我们将同一个核应用于采用不同编码方式（非独热编码）的不同序列。具体而言，设 $\tilde{\mathcal{B}}$ 是 $\mathbb{R}^{\mathcal{B}}$ 的另一个基。$\tilde{S} = \cup_{L=0}^{\infty}\tilde{\mathcal{B}}^L$ 中的每个序列 $\tilde{Y}$ 都可以表示为向量编码 $\tilde{V} \in \mathbb{R}^{|\tilde{Y}|\times\mathcal{B}}$，其中 $\tilde{V}_{(l)}$ 是字母 $Y_{(l)}$ 的编码 $\tilde{}$ $\tilde{}$。将 V 代入方程 4，我们定义 Y 的核嵌入；它由来自每个序列 $X \in S$ 的核嵌入 k_X 的线性组合构成，且满足 $\lrcorner|X| = |V|$。我们将证明，如果核 k 在 S 上具有离散质量，并将 $\tilde{}$ $\lrcorner k$ 应用于 S，它仍将具有离散质量。我们将字母表 $\mathcal{B}$ 到 $\mathcal{B}$ 的这种转换称为核的字母表“重新参数化”。

直观地说，在蛋白质的情形下，我们可以将重新参数化后的字母表视为一组氨基酸属性，例如质量、电荷等。每种氨基酸都是原始字母表 $\mathcal{B}$ 中的一个字母 ˜，而每种属性都是重新参数化字母表 $\mathcal{B}$ 中的一个字母；由于这两个字母表都是 $\mathbb{R}^{\mathcal{B}}$ 的基，因此每种氨基酸都可以表示为若干属性的线性组合。通过分析同一个核应用于氨基酸属性序列时的灵活性，我们可以分析该核在氨基酸序列上的灵活性。这在理论上有助于研究复杂的核。

命题 19. 设 k 是定义在 S上的严格正定核，且 $\tilde{\mathcal{B}}$ 是 $\mathbb{R}^{\mathcal{B}}$的一组基。˜那么， k 作为定义在 S 上的核具有离散质量，当且仅当它作为定义在 S上的核具有离散质量。

证明 $\mathcal{B}$ 和 $\tilde{\mathcal{B}}$ 都是 $\mathbb{R}^{\mathcal{B}}$的一组基，因此定义在 S 上的核是定义在 S 上的核的重新参数化˜，反之亦然。因此，我们只需证明：如果 k 作为定义在 S上的核具有离散˜质量，那么它作为定义在 S 上的核也具有离散质量。首先注意，对字母表进行重新参数化不会改变核嵌入向量张成空间的张成空间，$\mathrm{span}(k_V)_{V\in(\tilde{\mathcal{B}})^L} = \mathrm{span}(k_{V'})_{V'\in(\mathcal{B})^L}$，因为 $\mathcal{B}$和 $\tilde{\mathcal{B}}$ 都是 $\mathbb{R}^{\mathcal{B}}$的一组基。现在，考虑某个长度 L。如果 $X \in S$且 $|X| = L$，那么对某些 α_X和 $_Y$，有 $k_X = \sum_{Y\in\tilde{S}\ |\ |Y|=L} \alpha_{X,Y} k_Y$。由于 k是严格正定核， $\{k_X\}_{|X|=L}$是线性无关集，因此 α 是一个维度为 $|\mathcal{B}|^L \times |\mathcal{B}|^L$的可逆方阵。令$f_X = \sum_{Y\in\tilde{S}\ |\ |Y|=L} \left(\alpha^{-1}\right)_{X,}$ $_Y \delta_Y$。那么，对于满足 $|Z| \neq L$的 $Z \in S$，有 $f_X(Z) = 0$；并且如果 $|Z| = L$，则 $f_X(Z) = \sum_{Y\in\tilde{S}\ |\ |Y|=L} \left(\alpha^{-1}\right)_{X,}$ $_Y \alpha_{Z,Y} = \delta_X(Z)$。因此 $\delta_X \in \mathcal{H}_k$。 ■

在本文余下部分中，对于 S 上的任意核以及任意 V， $W \in \cup_{L=0}^{\infty}\mathbb{R}^{L\times\mathcal{B}}$ 我们将记作 $k(V, W) = (k_V|k_W)_k$，其中 k_V 和 k_W 由式 4 定义。

7. 逐位置比较核

本节中，我们设计具有离散质量的核，用于逐位比较序列。我们在第 3 节和第 5.2 节中看到，汉明核、加权度核以及其他相关的逐位比较序列核既不是通用核，也不具特征性。在这里，我们将开发其他核，它们既能体现相同的生物学思想，又具有很强的灵活性。

逐位序列比较在生物学中无处不在，并且对于许多问题都有充分的生物学依据。例如，一个常见的观察是，特定位置上的核苷酸或氨基酸具有特定的生物学功能；比如，若干特定位置上的氨基酸可能发生化学反应，形成荧光团，使蛋白质呈现荧光。因此，在根据序列预测表型时，一种合理的序列相似性测度是逐位比较序列，正如汉明核所做的那样。此外，进化过程中一种非常常见的突变形式是替换，即用一个字母替换另一个字母。因此，逐位序列相似性测度既能捕捉进化距离，也能捕捉表型距离；例如，在比较序列分布时，这可能是值得采用的做法。

我们的新核保留了现有的位置级序列相似性概念。例如，根据汉明核，两个序列之间的差异越来越大时，它们

根据我们提出的汉明核替代方案，结果也会越来越不同。我们修改的不是序列相似性的测度，而是核对序列相似性的依赖关系的泛函形式，即 X 与 Y 的相似性发生变化时，究竟如何转化为 $k(X,Y)$ 的变化。对于现有的核而言，其泛函形式很少有强有力的生物学依据，通常仅仅出于便利性的考虑。我们的结果表明，对于生物序列的核方法而言，泛函形式实际上对其可靠性有着极其重要的影响。

我们先研究一种简单的核，它逐位置比较序列。然后，我们使用这个“基础”核，借助第6.2节中开发的变换，推导出许多其他类型的逐位置比较核。基础逐位置比较核定义为：将分别作用于每个位置上字母的单个核相乘。

定义 20 (基础逐位置比较核)。 我们将每个序列 $X \in S$表示为以无限长的终止符号\$尾部结束。令 k_s 为定义在字母 $\mathcal{B} \cup \{\$\}$ 上的严格正定核，并满足 $k_s(\$,\$) = 1$。于是，基础逐位置比较核为

$$k(X,Y) = \prod_{l=0}^{\infty} k_s(X_{(l)}, Y_{(l)}). \tag{5}$$

注意，由于 $k_s(\$,\$) = 1$，无限乘积总是有限的。该核根据 k_s在每个位置 l比较序列 X 和 Y。例如， k_s 的一种自然选择是令 $k_s(b,b') = 1$ （若 $b = b'$ ），并令 $k_s(b,b') = e^{-\lambda}$ （若 $b \neq b'$），其中 $\lambda > 0$。由此得到“指数汉明核”，

$$k(X,Y) = \exp\left(-\lambda \sum_l \mathbb{1}(X_{(l)} \neq Y_{(l)})\right) = \exp(-\lambda d_H(X,Y)),$$

其中 d_H 是汉明距离。回想一下，滞后为 $L = 1$ 的汉明核形式为 $k(X,Y) = |X| \vee |Y| - d_H(X,Y)$。因此，这两个核都使用汉明距离来测度序列相似性，但它们依赖关系的泛函形式不同。

基础逐位置比较核不同于加权度核，具有离散质量。

定理 21。 基础逐位置比较核具有离散质量。

证明 我们的证明策略是重新参数化字母表 $\mathcal{B}$ （使用命题19），使得再生核希尔伯特空间 $\mathcal{H}_k$ 分解为一组正交超平面的乘积，每个正交超平面都由一个序列子集张成。我们证明，将核限制在这些正交超平面中的每一个上会呈现出简单的形式，并且具有离散质量。然后，我们应用命题 16 合并这些独立的正交超平面，并证明该结果。

我们首先对基础逐位置比较核 k 重新参数化。首先，根据 Schur 乘积定理，注意到 k 是一个严格正定核。将 K 定义为满足 $K_{b',b} = k_s(b,b')$（其中 $b,b' \in \mathcal{B}$）的矩阵。将 $t \in \mathbb{R}^{\mathcal{B}}$ 称为满足 $t_b = k_s(b,\$)$（其中 $b \in \mathcal{B}$）的向量。令 $U = K^{-1}t$，并将其称为 $\sigma = U^T K U$。如果 $V, W \in \mathbb{R}^{\mathcal{B}}$，则类似地定义 $k_{s,V} = \sum_{b\in\mathcal{B}} V_b k_{s,b}$，以及 $k_{s\,W}$ $k_{s\,V\,W} = k_{s\,V}|k_{s\,W}\ k_s$ $k_{s\,V} = k_{s\,V}|k_{s\,\$}\ k_s$ $V \in \mathbb{R}^{\mathcal{B}}$，(,) (,,) 和 (,\$) (,,) 。那么，如果

$k_s(V,U) = V^T K U = V^T t = k_s(V,\$)$。特别注意，由于 k_s 是严格正定的，我们有严格的柯西-施瓦茨不等式，

$$\sigma = k_s(U,U) = k_s(U,\$) < \sqrt{k_s(U,U)k_s(\$,\$)} = \sqrt{\sigma},$$

这只有在 $\sigma < 1$ 时才可能成立。令 B_1、.....、$B_{|\mathcal{B}|-1} \in \mathbb{R}^{\mathcal{B}}$ 满足条件，使得在使用点积 $(v|x) = v^T K w$，且对所有 i 均有 $(B_i|B_i) = 1$ 时，$\mathbb{R}^{\mathcal{B}}$ 是向量空间的正交基。因此，对所有 i、j，$k_s(B_i, B_j) = \delta_i(j)$，并且对所有 i，~ $k_s(U, B_i) = 0$。将 $\mathcal{B} = \{U, B_1, \ldots, B_{|\mathcal{B}|-1}\}$ 和 S 称为由 $\mathcal{B}$ 中的~ ~个字母组成的序列集合。根据命题 19，如果我们能证明核 k 在 S 上具有离散质量，就能知道它在 S 上也具有离散质量。

~现在我们把 S 拆分为若干独立的定义域 W_V。回顾一下，基础的位置逐项比较核形式为 $k(V, V') = \prod_{l=0}^{\infty} k_s(V_l, V_l')$。应用于重新参数化字母表中的序列时，即 $V, V' \in S$，对所有 $b, b' \in \mathcal{B} \cup \{\$\}$，有 $k_s(b, b') = 0$，当且仅当 $b \neq b'$，但 $k_s(U, \$) = \sigma$ 除外。因此，如果对 $X, Y \in S$ 有 $|X| \geq |Y| = L$，那么 $k(X,Y) \neq 0$ 当且仅当 X 和 Y 的前 L 个字母相同，且 X 在位置 L 之后的所有字母均为 U。若 $V \in S$ 且不以 U 结尾，则将 $W_V = \{V + m \times U\}_{m=0}^{\infty}$ 定义为 S 中所有以 V 开头且尾部由 U 个 s 组成的序列~ ~。那么，如果 $V' \in S$ 是另一个不以 U 结尾的不同序列，则 $W_{V'}$ 与 W_V 在 $\mathcal{H}_k$ 中正交。因此，$\mathcal{H}_k$ 由集合 $\{W_V\}_V$ 张成的正交超平面构成。根据命题 16，k 具有离散质量，当且仅当 k 限制在每个 W_V 上时具有离散质量。

现在我们证明，应用于每个 W_V 的核都具有离散质量。注意，由于 $k_s(U, \$) = k_s(U, U) = \sigma$，我们有 $k(V + m \times U, V + m' \times U) = k(V,V) k(m \times U, m' \times U) = k(V,V)\sigma^{m \vee m'}$。因此，$k$ 限制在 W_V 上等价于核 $k^{(n)}(m, m') = \sigma^{m \vee m'}$ 应用于 $m, m' \in \mathbb{N}$，再乘以常数 $k(V,V)$。因此，我们只需证明 $k^{(n)}(m, m')$ 这个定义在自然数上的一维核具有离散质量。我们用归纳法证明这一点。首先，注意到 $\sigma < 1$，令 $f = (\sigma - 1)^{-1}(k_1^{(n)} - k_0^{(n)})$。可见 $\delta_0 = f$，因此 $\delta_0 \in \mathcal{H}_k$。现在假设 $\delta_0, \ldots, \delta_{x-1} \in \mathcal{H}_k$ 对某个 $x \in \mathbb{N}$ 成立。令 $f_x = (\sigma - 1)^{-1}(k_{x+1}^{(n)} - k_x^{(n)})$。可见，如果 $z > x$，则 $f_x(z) = 0$；并且 $f_x(x) = 1$。因此 $\delta_x = f_x - \sum_{z=0}^{x-1} f(z)\delta_z$，所以 $\delta_x \in \mathcal{H}_k$。根据命题 16，证明完成。 ■

我们可以利用基础的逐位置比较核来构造更复杂的逐位置比较核，以应对特定的建模挑战，同时还具有离散质量。在下面每个示例中，为简单起见，我们使用 $k_s(b, b') = \exp(-\lambda \mathbb{1}(b \neq b'))$，但同样也可以使用任意严格正定核 k_s，例如基于氨基酸相似性比较 b 和 b' 的核。

示例 6（重尾汉明核）。实际挑战：指数汉明核具有“轻尾”特性，即随着汉明距离增加而快速（呈指数级）衰减。考虑一个由 $\{X_1, X_2, \ldots\}$ 远距离相关序列组成的数据集，以及其格拉姆矩阵 K，其元素为 $K_{ij} = k(X_i, X_j)$。轻尾可能导致“对角占优”，即对角元素 K_{ii} 远大于非对角元素 K_{ij}，对于 $i \neq j$ 而言。因此，该核的行为可能非常类似于单位核（示例 4），并且在高斯过程或其他回归方法中使用时表现出同样较差的泛化能力。通过调节 λ 来解决对角占优问题很困难；通常唯一可用的其他状态是矩阵 K 几乎为常数，即

$K_{i,\ j} \approx constant$ 对于所有 i、j。在这种情况下，回归将在各处做出近似相同的预测，并且仍会表现出较差的泛化能力。

提出的核：为了解决对角占优问题，我们提出了一种重尾核，其值会随着序列距离的增加而缓慢衰减。设 $\beta, C > 0$。则"逆多二次汉明核"通过对带宽参数 λ 积分来定义，

$$k(X,Y) = (C + d_H(X,Y))^{-\beta} = \Gamma(\beta)^{-1} \int_0^\infty d\lambda \left(\lambda^\beta e^{-C\lambda}\right) e^{-\lambda d_H(X,Y)}, \tag{6}$$

其中 Γ 是伽马函数。随着汉明距离增加，该核按幂律而非指数规律递减，即其尾部非常厚。根据命题 ***14***，该核具有离散质量。

示例 7 (k-mer 汉明核)。 实际挑战：指数汉明核认为，如果两个序列在相同位置具有相同的字母，那么它们就是相似的。然而，序列的生物学功能不仅取决于某个特定位置是否有某个字母，还可能取决于某个特定位置是否有一串特定的字母（例如一个基序）。因此，根据两个序列在相同位置是否具有相同的 ***k-mer***，而非单个字母，来判断它们的相似性或许更合适。这就是滞后 $L > 1$ 的汉明核背后的思想。问题在于，该核既不具特征性，也不具通用性（第 ***3*** 节）。

提出的核 我们提出一种核，它采用与滞后为 L 的汉明核相同的序列相似性度量，但采用不同的泛函形式，从而确保其具有离散质量。设 L 为上下文大小，并定义 $\tilde{\mathcal{B}} = \mathcal{B}^L$ 为长度为 L 的 ˜ 序列集合。我们将 $\mathcal{B}$ 视为字母表，并考虑由该字母表构成的序列集合 $\tilde{S}$。现在考虑定义在 $\tilde{S}$ 上的核 $\tilde{k}(X,Y) = (C + d_H(X,Y))^{-\beta}$，其中 C、$\beta > 0$ 为任意值；回想一下，该核具有离散质量。 S 中的每个序列（即原始序列空间）˜ ˜ 都可以通过映射 $i: S \to \tilde{S}$ 唯一地嵌入 $\tilde{S}$，该映射定义为

$$h(X) = (X_{(:L)}, X_{(1:L+1)}, X_{(2:L+2)}, \dots),$$

也就是说， h 按顺序将每个序列映射为其 L-元组序列。注意， h 是单射，但 ˜ 不是满射。根据命题 ***16***，将 k 限制在 h 的像上后，它具有离散质量。因此，"滞后为 L 的逆二次汉明核"，

$$k(X,Y) = \tilde{k}(h(X), h(Y)) = \left(C + \sum_{l=0}^{|X|\vee|Y|-1} \mathbb{1}(X_{(l:l+L)} \neq Y_{(l:l+L)})\right)^{-\beta},$$

该函数为 X、$Y \in \tilde{S}$ 定义，具有离散质量。请注意，上述表达式中的求和恰好是滞后 L 的汉明核。换言之，序列相似性的概念并未改变，但核的泛函形式发生了变化，这使得新的核具有离散质量。

示例 8 （中心对齐的汉明核）。 实际挑战 指数汉明核从位置一开始，逐位比较序列。当位置一是一个良好的参照点（例如基因的起始位置）时，这种做法很合理。然而，在某些情况下，我们感兴趣的是参照点两侧的某个序列区域，

例如起始密码子上游和下游的 *DNA*序列（即启动子区和编码区）。例如，与其按如下方式比较序列，

$$|AACTTCT\$\$\$\$...$$
$$|GGACTTCTCT\$...$$

其中 | 标记参考点，我们希望按如下方式对它们进行比较，

$$...\$\$AACT|TCT\$\$\$...$$
$$...\$GGACT|TCTCT\$...$$

请注意，现在不同序列在参考点 |两侧的长度可能各不相同。

提出的核 我们可以将每个数据点理解为由一对序列而非单个序列构成：$(X^{(1)}, X^{(2)})$，其中 $X^{(1)}$ 是参考点左侧的序列， $X^{(2)}$ 是右侧的序列。因此，每个数据点属于 S^2 ，而不是 S。从指数汉明核（或任何具有离散质量的其他核）出发，我们可以利用张量化将其扩展为 S^2 ，而命题 *17* 保证它具有离散质量。

示例 9 (带偏移的汉明核**)**。 实际挑战 两个序列之间究竟应该比较哪些位置，并不总是显而易见。例如，蛋白质序列通常有明确的起始位置（起始密码子），而对于其他遗传元件（如增强子），序列的"开端"则不那么明确。解决这一问题的一种方法是像带偏移的汉明核那样，在彼此之间采用不同的偏移量来比较序列（*Sonnenburg* 等，*2007*年）。然而，带偏移的汉明核并不具有离散质量。

提出的核 我们可以将移位的逐位置比较核定义为

$$k_L(X,Y) = \sum_{l=0}^{L} k(X_{(l:)}, Y) + k(X, Y_{(l:)}),$$

其中， k 例如是指数汉明核。该核在各种偏移下将 X 与 Y进行比较，即从位置 $l = 0$开始，然后是位置 $l = 1$，以此类推，直到某个最大值 L；如果这些偏移中至少有一个能产生良好匹配，该核的值就会很大。该核是具有离散质量的各个核之和，因此根据命题 *14*，它具有离散质量。

请注意，这些提出的核除了具有离散质量之外，在计算上也是易于处理的，尤其可以使用为加权度核开发的技术进行高效计算（Sonnenburg 等，2007年）。

8. 比对核

在本节中，我们设计具有离散质量的核，依据成对比对来比较序列。逐位置比较核只有在少量替换即可将 X 转变为 Y 时，才会认为两个序列 X 和 Y 相似；而比对核则认为，只要少量

替换、插入和/或删除即可将 X 转变为 Y，两个序列便相似。从生物学角度看，以这种方式相似的序列通常具有相似的表型，且在进化上关系密切。然而，逐位置比较核难以充分捕捉这种相似性。例如，序列 $X = ATGC$ 与序列 $Y = TGC$ 的差异仅在于前者在序列开头插入了一个 A，但基于汉明距离的核会认为这两个序列完全不同，其距离为 $d_H(X, Y) = 4 = |X|$。

处理插入和缺失的一种常见但存在问题的技术，是使用多序列比对算法对数据集进行预处理；该算法会在每条序列中添加空位符号，使其长度相同。从本质上说，多序列比对算法旨在对各序列中哪些字母通过替换而在进化上彼此相关作出点估计，并将每个相关字母置于相同位置，即预处理后数据矩阵的同一列中。序列完成比对后，应用逐位置比较核就更为合理；通过将空位符号纳入字母表 $\mathcal{B}$，即可比较插入和缺失。多序列比对预处理的问题在于：(a) 它没有考虑哪些位置彼此相关这一判断中的不确定性，而是依赖点估计；(b) 它会阻止下游机器学习方法泛化到未见过的序列，因为向数据集添加新序列可能会改变多序列比对，例如，新序列比此前观察到的序列更长时就会如此（Weinstein and Marks，2021年）。

比对核为处理插入和缺失提供了另一种方法（Haussler，1999）。它们不转换整个数据集，而是认为两个序列 X 和 Y 相似，只要它们之间的差异仅涉及少量插入和缺失以及替换。比对核对于涉及长度变化较大的序列的问题尤为重要，例如分析人类抗体或关系较远的进化同源物。

8.1 比对核

本节定义比对核，并说明当且仅当其超参数设为某个范围内的值时，它才具有离散质量。

在生物序列分析中，两个序列之间的成对比对是指分别从每个序列中选取一个位置子集并进行匹配，同时满足以下限制：（a）每个序列中的每个位置最多只能与另一个序列中的一个位置匹配；（b）匹配必须按顺序排列，因此，如果序列 X 中的位点 l_X 与序列 Y中的 l_Y 匹配，那么 X 中的位点 $l'_X > l_X$ 就不能与 Y 中的位点 $l'_Y < l_Y$ 匹配（Durbin 等人，1998，第 2 章）。例如，序列 X =ATGC 和 Y =ACC 的一种比对是，

```
ATGC
|| |
AC-C
```

其中竖线 | 表示匹配。这里，- 是空位符号，表示核苷酸 G 在 X 中未匹配；我们可以将 G 解释为相对于Y而言在 X 中的一次插入；反过来，也可以将空位 - 解释为相对于 X而言在 Y中的一次删除。比对核

$k(X,\ Y)$ 会考虑 X 与 Y之间所有可能的成对比对，对每种比对应用逐位置比较核，对对齐序列进行评分，并将这些得分相加作为其值。

从数学上说，要定义比对核，我们先从两个较简单的核入手，然后按照 Haussler（1999）的构造方法将它们卷积。第一个核 k_s 用于评估比对字母之间的匹配。第二个核 k_I 用于评估插入和删除；它采用 “ 仿射 ” 间隙罚分，分别对间隙的出现及其总长度进行惩罚。对于定义在 S 上的任意两个核 k_1 和 k_2，卷积核 $k_1 \star k_2$ 为：

$$(k_1 \star k_2)(X,Y) = \sum_{X^{(1)}+X^{(2)}=X} \sum_{Y^{(1)}+Y^{(2)}=Y} k_1(X^{(1)},Y^{(1)})k_2(X^{(2)},Y^{(2)})$$

其中，求和遍历所有满足 $X^{(1)} + X^{(2)} = X$ 且 $Y_{(1)}+Y_{(2)} = Y$ 的序列 $X^{(1)}, X^{(2)}, Y_{(1)}Y_{(2)} \in S$。请注意，$S$ 包含长度为零的序列，因此求和也包含诸如 $X^{(1)} = X, X^{(2)} = \emptyset$ 这样的项。

定义 22（比对核）。设 k'_s 是 $\mathcal{B}$上的一个严格正定核。将 k'_s 扩展到 S，并令k_s 满足：当 $|X| \neq 1$ 或 $|Y| \neq 1$时，$k_s(X,Y) = 0$。

令 $\Delta\mu > 0$ 为开始一次插入的惩罚，令 $\mu > 0$ 为插入长度的惩罚。插入惩罚函数 g 定义为：若$|X| = 0$，则为 $g(X) = 1$；否则为 $g(X) = \exp(-\Delta\mu - |X|\mu)$。对 $X, Y \in S$，定义半正定核$k_I(X,Y) = g(X)\,g(Y)$。

比对核 k 为，

$$k = \sum_{l=0}^{\infty} k_I \star (k_s \star k_I)^{\star l},$$

其中，指数 $\star l$ 表示 l 个核 $k_s \star k_I$的卷积。

为说明该核确实考虑了 $X,\ Y \in S$的所有可能成对比对，请注意，我们可以将 $k(X,\ Y)$ 重写为：

$$\sum_{l=0}^{\infty} \sum_{X^{(1)}+\cdots+X^{(2l+1)}=X} \sum_{Y^{(1)}+\cdots+Y^{(2l+1)}=Y} k_I(X^{(1)},Y^{(1)}) \prod_{i=1}^{l} k_s(X^{(2i)},Y^{(2i)})k_I(X^{(2i+1)},Y^{(2i+1)}). \tag{7}$$

求和中的每一项都表示一种比对。每种比对都有 l 个匹配，其中偶数子序列 $X^{(2)}, X^{(4)}, \ldots, X^{(2l)}$（属于 X）与偶数子序列 $Y(2)\,Y(4)\ldots, Y(2l)$（属于 Y）相匹配。要使该项非零，这些子序列都必须只包含一个字母，否则 $k_s(X^{(2i)}, Y(2i)) = 0$。在 X 和 Y 中的每对匹配位置之间都可以有插入，对应于奇数子序列$X^{(1)}, X^{(3)}, \ldots, X^{(2l+1)}$和 $Y(1)\,Y(3)\ldots, Y(2l+1)$。这些子序列的长度可以任意，包括零。本节前面看到的 $X = \mathsf{ATGC}$ 与 $Y = \mathsf{ACC}$ 之间的比对示例对应于 $X^{(1)} = \emptyset$, $X^{(2)} = \mathsf{A}$, $X^{(3)} = \emptyset$, $X^{(4)} = \mathsf{T}$, $X^{(5)} = \mathsf{G}$, $X^{(6)} = \mathsf{C}$,$X^{(7)} = \emptyset$, $Y(1) = \emptyset$, $Y(2) = \mathsf{A}$, $Y(3) = \emptyset$, $Y(4) = \mathsf{C}$, $Y^{(5)} = \emptyset$, $Y^{(6)} = \mathsf{C}$, $Y(7) = \emptyset$。

该核对所有可能的比对求和：公式 7 中最外层的求和遍历匹配数 $l \in \{0, 1, 2, \ldots\}$，内层求和则遍历 X 和 Y 中所有可能的匹配位置选择。然后，根据 k_s 和 k_I 为每种比对计算得分。

比对核的灵活性如何？Haussler（1999）证明了它是严格正定的。Jorgensen和Tian（2015年）受生物序列分析之外问题的启发，证明了“二项式再生核希尔伯特空间”没有离散质量；这对应于满足 $\Delta\mu = \mu = 0$、$|\mathcal{B}| = 1$ 和 $k_s(b, b) = 1$ 的比对核。现在我们证明，当且仅当 μ，$\Delta\mu$ 和 k_s 满足某些约束时，比对核才可能具有离散质量。

定理 23（比对核可以具有离散质量）。定义 K 为矩阵，其中 $K_{b,\,b'} = k_s(b, b')$，对于 b、$b' \in \mathcal{B}$ 和 $\sigma = \mathbf{1}_{\mathcal{B}}^T K^{-1} \mathbf{1}_{\mathcal{B}}$，其中 $\mathbf{1}_{\mathcal{B}} \in \mathbb{R}^{\mathcal{B}}$ 是每个元素均为1 的向量。如果 $\Delta\mu > 0$，则 k 具有离散质量，当且仅当 $2\mu \geq \log\sigma$。如果 $\Delta\mu = 0$，则 k 具有离散质量，当且仅当 $2\mu > \log\sigma$。

证明见附录 E。它采用了与证明逐位置比较核具有离散质量（定理 21）时类似的策略：对字母表重新参数化，然后将序列空间分解为正交超平面。在这些正交超平面上，核具有更简单的形式；我们利用比对核的特征表示证明该形式具有离散质量。相关特征将在下文第 9.1 节推导。

从实际应用来看，定理 23 表明，如果我们不加斟酌地设置比对核的超参数，它可能变得不可靠；而如果设置得当，则能保证其灵活性。特别是，如果我们希望具有灵活性，缺口惩罚（μ 和 $\Delta\mu$）的选择会受到替换惩罚（k_s）的约束，具体关系为 σ。如果开始缺口不受惩罚（$\Delta\mu = 0$），则为了使核具有离散质量，延长缺口的惩罚（μ）必须大于$_1{}_2\log(\sigma)$。如果开始缺口需要惩罚（$\Delta\mu > 0$），延长缺口的惩罚可以略小一些，因为 $\mu = \frac{1}{2}\log(\sigma)$ 同样能使核具有离散质量。在附录 F 中，我们说明这些条件源于以下要求：应用于单个比对的逐位置比较核具有离散质量。在附录 H 中，我们探讨了对这些条件的放宽，使比对核不再具有离散质量，但仍然是通用的；我们发现，只有对核进行额外且有问题的修改，这些放宽条件才成立。

8.2 重尾对齐核

实际上，当应用于具有高度多样化序列的数据集时，使用比对核的回归方法通常会因对角占优而表现出较差的泛化能力（Haus- sler，1999；Saigo 等人，2004；Weston 等人，2003）。以往解决这一问题的尝试遇到了相当大的困难；实践者甚至引入了并非半正定的替代“核”（Saigo 等人，2004）。本节将诊断比对核中对角占优的可能来源，并提出比对核的重尾修改形式，同时保留离散质量。

关键观察是，比对核实际上将逐位置比较核应用于序列之间的每一组成对比对；对角占优源于这一逐位置比较核的尾部较薄。更准确地说，

考虑一个比对核（定义 22），其字母核为 $k'_s(b, b') = \exp(-\lambda \mathbb{1}(b \neq b'))$; 回想一下，这一选择在第7节中为我们得到了指数汉明核。定义比对核的求和式（公式 7）中的每一项都对应一个独立的比对，并包含一个因子，该因子对应于将指数汉明核应用于匹配位置，即 $\prod_{i=1}^{l} k_s(X^{(2i)}, Y^{(2i)})$。随着匹配位置之间的汉明距离增大，该项会迅速（呈指数级）衰减，从而导致对角占优。为解决这一问题，我们沿用例 6 中的相同策略，对带宽参数 λ 进行积分，以推导出一种厚尾替代方案。

例10 (匹配上的重尾比对核)。“匹配上的厚尾比对核”是核 $\tilde{k}(X, Y) = \Gamma(\beta)^{-1} \int_0^\infty d\lambda \left(\lambda^\beta e^{-C\lambda}\right) k_\lambda(X, Y)$，其中 k_λ 是使用参数为 λ 的 k'_s 的比对核，且 $C, \beta > 0$。根据公式 *6*，我们发现 $\tilde{k}$，$k(X, Y)$ 为，

$$\sum_{l=0}^{\infty} \sum_{X^{(1)}+\cdots+X^{(2l+1)}=X} \sum_{Y^{(1)}+\cdots+Y^{(2l+1)}=Y} (C + d_H(X^{(2)} + \cdots + X^{(2l)}, Y^{(2)} + \cdots + Y^{(2l)}))^{-\beta} \times k_I(X^{(1)}, Y^{(1)}) \prod_{i=1}^{l} k_I(X^{(2i+1)}, Y^{(2i+1)})$$

在这个核中，匹配位置之间汉明距离的依赖关系遵循幂律，而非指数衰减。若定义 $K_{\lambda,\ b,\ b'} = k_\lambda(b, b')$，则 $\sigma_\lambda = \mathbf{1}_{\mathcal{B}}^T K_\lambda^{-1} \mathbf{1}_{\mathcal{B}} = \left(\frac{1}{|\mathcal{B}|} + \frac{|\mathcal{B}|-1}{|\mathcal{B}|} e^{-\lambda}\right)^{-1}$，当 $\lambda \to 0$ 时趋近于1 。因此，只要 $\mu > 0$, $\tilde{k_\lambda}$，根据定理 *23*，对于较小的 λ 就会具有离散质量。因此，根据命题 *14*，k 具有离散质量。请注意，这种改进后的核与标准比对核一样，可以使用动态规划算法高效计算（附录 *K*）。

除了替换惩罚 λ之外，我们还可以考虑插入惩罚 μ 和 $\Delta\mu$的作用。我们发现了类似的问题：随着插入长度增加，比对核的每一项都会呈指数衰减，因为 $k_I(X, Y) = \exp\left(-\mathbb{1}(|X| > 0)(\Delta\mu + \mu|X|) - \mathbb{1}(|Y| > 0)(\Delta\mu + \mu|Y|)\right)$。现在，参数 μ 起到了带宽的作用。为了产生厚尾，我们可以进行类似的变换，并取积分 $\Gamma(\beta)^{-1} \int_0^\infty d\mu \left(\mu^\beta e^{-C\mu}\right) k_\mu(X, Y)$，其中 k_μ 是参数为μ的比对核。由此得到一个间隙上的重尾比对核，根据命题 14，它具有离散质量。我们甚至可以对匹配上的厚尾比对核应用相同的变换，从而得到一个同时作用于间隙和匹配的厚尾比对核。

8.3 Local alignments

基因组非常长，因此在实践中，机器学习通常只针对特定的遗传元件进行，例如基因、启动子、增强子等。这些遗传元件究竟从哪里开始或在哪里结束，有时并不明确。这促使人们提出局部比对的概念，它忽略序列开头和结尾处的插入和缺失。使用局部比对作为相似性的测度，意味着包含相同遗传元件但侧翼区域不同的两个序列仍会被视为相似（Durbin 等人，1998，第 2章）。在本节中，我们研究局部比对核，并建立其具有离散质量的条件（Vert 等人，2004年；Saigo 等人，2004年）。

局部比对核是比对核的一种修改版本，不会惩罚在成对比对的开头或结尾创建间隙。

定义 24 (局部比对核**)**。 设 k_I 和 k_s 如定义 *22* 中所述。另定义˜修正后的插入间隙惩罚核 $k_I(X,Y) = \exp(-\mu(|X|+|Y|))$，该核不含间隙起始惩罚 $\Delta\mu$。局部比对核为，

$$k_{\mathrm{la}} = \tilde{k}_I + \sum_{l=1}^{\infty} \tilde{k}_I \star (k_s \star k_I)^{\star(l-1)} \star k_s \star \tilde{k}_I.$$

我们可以扩展定理 23 证明中的思路，得出局部比对核具有离散质量的条件；结果表明，这些条件与常规对齐核的条件完全相同。

定理 25。 定义 K 为矩阵，其元素为 $K_{b,b'} = k_s(b,b')$，其中 $b, b' \in \mathcal{B}$，并且 $\sigma = \mathbf{1}_{\mathcal{B}}^T K^{-1} \mathbf{1}_{\mathcal{B}}$，其中 $\mathbf{1}_{\mathcal{B}} \in \mathbb{R}^{\mathcal{B}}$ 是每个分量均为 1 的向量。若 $\Delta\mu = 0$, k_{la}，则其具有离散质量当且仅当 $2\mu > \log\sigma$。若 $\infty > \Delta\mu > 0$, k_{la}，则其具有离散质量当且仅当 $2\mu \geq \log\,\sigma$。若$\Delta\mu = \infty$, k_{la}，则无论 μ, σ取何值，它都具有离散质量。

证明见附录 J。请注意，对于标准非局部比对核，如果我们将间隙起始惩罚设为无穷大（$\Delta\mu = \infty$），就会得到逐位置比较核。然而，对于局部比对，情况更复杂，因为每个序列开头和结尾的间隙数量有多种选择。我们发现，当 $\Delta\mu = \infty$ 时，局部比对核具有离散质量，而与其其他超参数的取值无关。

9. k-mer 谱核

在本节中，我们研究 k-mer谱核。k-mer 谱核与比对核一样，其动机源于这样的生物学观察：因插入、缺失或其他复杂突变而不同的序列，通常具有相似的表型，并且在进化上彼此相关。值得注意的是，我们将发现，k-mer谱核与比对核之间的关系比它们的动机更为深刻：在某些情况下，二者在适当的倾斜下是等价的。

k-mer 谱核根据子序列（K-mer）的存在与缺失对序列进行特征化。由于 K-mer 在序列中的位置不会影响相应特征的值，与逐位置比较核相比，k-mer 谱核可能对插入或删除相对不敏感。在例1中，我们看到了一个常用的 k-mer 谱核，它只有有限数量的特征，因此既不是通用的，也不是特征性的。在这里，我们将研究无限维的 k-mer 谱核，并说明它们具有离散质量。

9.1 带间隙 k-mer 谱核

我们首先将比对核与 k-mer 谱核联系起来。我们证明，对于特定的倾斜方式，比对核的特征是 k-mer 计数的加权和，并允许存在一定的 k-mer 错配。更准确地说，我们证明倾斜比对

核等价于无限维间隔 k-mer 谱核。由于倾斜保留离散质量（命题 18），该间隔 k-mer 谱核具有离散质量。

直观地说，“间隔 k-mer”是包含间隙的 k-mer；如果另一个序列中的所有非间隔字符都匹配，而间隔位置上的字母无论是什么都不影响匹配，则两者匹配。例如，间隔 k-mer 序列 AT-G 与 ATCG、ATGG、ATAG 和 ATTG 匹配。形式上，长度为 L、包含 M 个非间隔字母的间隔 k-mer，由一个有序索引集$J = (j_0, \ldots, j_{M+1})$ 和一个由M个非间隔字母组成的序列 $Z \in \mathcal{B}^M$ 定义，其中 $-1 = j_{-1} < j_0 < j_2 < \cdots < j_M < j_M = L$。每个 j_m（$m \in \{0, \ldots, M-1\}$）索引都表示非间隔字母 $Z_{(m)}$ 的位置。如果 $j_m + 1 \neq j_{m+1}$，则第 m 个字母后有一个间隙。令 $G(L, M, g)$ 表示所有长度为 L、包含 M 个非间隔字母且有 g 个间隙的间隔 k-mer 索引 J 的集合。令$G(L, g) = \cup_M G(L, M, g)$ 表示对所有可能的 M 取并集。对于任意序列 $X \in S$ 和$J = (j_0, \ldots, j_{M+1}) \in G(|X|, g)$，定义 $X_{(J)} = X_{(j_0)} + \cdots + X_{(j_{M-1})}$。

下面的定理说明，比对核等价于带间隔 k-mer 频谱核。具体而言，该定理表明，我们可以用特征 $u_V(X)$ 来表示比对核；这些特征取决于序列 $V \in S$ 在 X 中出现的次数，并允许存在间隙。换言之，每个特征 $u_V(X)$ 都取决于对所有非间隔字母为 V 的带间隔 k-mer 计数求和。

定理 *26*（比对核的特征是带间隔 *k-mer* 计数）。考虑一个比对核 k，其满足 $k_s(b, b') = \sigma^{-1}|\mathcal{B}|\delta_{b'}(b)$，其中 $b, b' \in \mathcal{B}$。将核按 $A(X) = e^{\mu|X|}$ 倾斜后，得到 $k^A(X, Y) = A(X)A(Y)k(X, Y)$。对于所有序列 $X \in S$，定义特征$\{u_V(X)\}_{V \in S}$如下：

$$u_V(X) = e^{\frac{1}{2}\zeta|V|} \sum_{g=0}^{\infty} e^{-\Delta\mu g} \sum_{J \in G(|X|, g)} \mathbb{1}(X_{(J)} = V), \tag{8}$$

其中 $\zeta = 2\mu -$ 对数 $\sigma +$ 对数 $|\mathcal{B}|$。现在，$\{u_V(X)\}_{V \in S}$ 是 $\mathcal{H}_{k^A}$ 的标准正交基。我们将核 $k^A(X, Y)$ 称为“带间隔 *k-mer* 频谱核”。

证明见附录 G。该定理表明，倾斜比对核的特征取决于带间隔 k-mer 计数的加权和。随着间隙数量增加，权重 $e^{-\Delta\mu g}$ 会减小，从而对间隙数量较多的带间隔 k-mer 施加惩罚。由于倾斜保留了离散质量，我们可以得出结论：当且仅当满足定理 23 的条件时，带间隔 k-mer 频谱核具有离散质量。

定理 26 中的倾斜会降低较长序列的权重，但不会改变比对核评估序列相似性的方式。实际上，该定理很容易推广到其他倾斜，例如常用的归一化倾斜 $A(X) = \sqrt{k(X, X)}^{-1}$。在这种情况下，我们得到 $k^A(X, Y) = {}^{u(X)^T u(Y)}/_{\|u(X)\|\|u(Y)\|}$，其中 $u(X)$ 是一个以 S 为索引的无限维向量，其第 $V \in S$ 个位置的值为 $u_V(X)$。该定理还可以推广到非对角 k_s，此时特征是基于重新参数化字母表的带间隔 k-mer 计数（见附录 G）。

定理 26 提供了一个带有离散质量的 k-mer 谱核的实用示例，也可作为理论工具。定理 45（附录 I）利用定理 26 证明，简单版本的比对核具有离散质量，进而使我们能够

证明完整比对核具有离散质量（定理 23）。定理 45 所用的证明技巧，依赖于对生物序列比对组合结构的细致分析，并运用了生成函数技巧和 Riordan 群理论（Shapiro 等，1991）。

9.2 重尾间隔 k-mer 谱核

现在，结合定理 26，我们重新考察比对核中的对角占优问题。公式 8 表明，特征权重，即 $\exp(\frac{1}{2}\zeta|V|)$，会随 k-mer 长度 $|V|$ 指数增长。这种指数级缩放意味着，当核 $k(X,\ Y)$ 用于非常相似的序列 X 和 Y时，其取值会非常大；相比之下，当X 和 Y 差异很大、仅有较短的共同 k-mer 时，其取值则小得多。这可能导致对角占优。试图通过令 ζ 变小来避免这个问题，会导致灵活性下降，因为当 $\zeta < \log|\mathcal{B}|$时，该核不再具有离散质量。换句话说，$\zeta$ 是一个调节参数，用于权衡核灵活性与对角占优。为了解决这个问题，我们提出一种具有离散质量的间隔 k-mer 核，其中特征权重随 k-mer 长度按幂律而非指数规律缩放。

例 11 (重尾带间隔 k-mer 频谱核**)**。 将未缩放的特征称为$\tilde{u}_V(X) = e^{-|V|\zeta/2}u_V(X)$，并注意到它们不依赖于 ζ。考虑一个由 $\underline{A}(X) = e^{-\frac{1}{2}\zeta|X|}$ 倾斜的带间隔 ***k-mer*** 谱核，

$$k_\zeta(X,Y) = e^{-\frac{1}{2}\zeta|X|}e^{-\frac{1}{2}\zeta|Y|}\sum_{V\in S} u_V(X)u_V(Y) = \sum_{V\in S} e^{-\frac{1}{2}(|X|+|Y|)+|V|}\tilde{u}_V(X)\tilde{u}_V(Y).$$

那么对于 C、$\beta > 0$，“重尾带间隔 ***k-mer*** 频谱核”定义如下：

$$\begin{aligned}\tilde{k}(X,Y) =& \Gamma(\beta)^{-1}\int_0^\infty k_\zeta(X,Y)\left(\zeta^\beta e^{-C\zeta}\right)d\zeta \\ =& \sum_{V\in S}\left(C+\frac{1}{2}(|X|+|Y|)-|V|\right)^{-\beta}\tilde{u}_V(X)\tilde{u}_V(Y).\end{aligned}$$

在此核中，较短的特征 V 按照幂律而非指数衰减进行降权。命题 *14*、定理 *23* 和定理 *26* 共同保证该核具有离散质量。该核可以使用附录 *K* 中给出的动态规划算法高效计算。

请注意，由于 ζ 取决于 μ，该核与第 8.2 节讨论的间隙上的重尾比对核密切相关；后者使用相同的基准测度对 μ 积分。成对比对得分随间隙长度增加而指数衰减，以及较短 k-mer 的特征权重呈指数衰减，是观察比对核中同一对角占优问题的两种方式。

9.3 无间隔 k-mer 谱核

现在我们考虑无间隔 k-mer 谱核，其特征对应于某个 k-mer 在序列中精确出现的次数。具体而言，我们证明，具有无限间隙起始惩罚 $\Delta\mu = \infty$ 的局部比对核是无间隔 k-mer 谱核，因此该 k-mer 谱核具有离散质量。

命题 27 （无限k-mer 谱核）。考虑一个倾斜局部比对核$k = k_{\text{la}}^{A}$，其具有 $\Delta\mu = \infty$、$\zeta = 0$ 和 $A(X) = \exp(\mu|X|)$。该核具有离散质量，并且可以写成

$$k(X,Y) = \sum_{V \in S} u_{\text{la},V}(X) u_{\text{la},V}(Y),$$

其中，特征 $u_{\text{la},\,V}(X) = \sum_{l=0}^{|X|-|V|} \mathbb{1}(X_{(l:(l+|V|))} = V)$ 表示 ***k-mer*** V 在 X中出现的次数。

证明 定义 $G_{\text{la}}(L,\ M,\ g)$ 为 $G(L) = \cup_g G(L,\ g)$ 中长度为 M且含有 g 个间隙的元素，不计开头或结尾处的间隙。按照定理 26 的相同思路，倾斜后，核的特征为：对于 $V,\ X \in S$

$$u_{\text{la},V}(X) = \sum_{g=0}^{\infty} e^{-\Delta\mu g} \sum_{J \in G_{\text{la}}(|X|,g)} \mathbb{1}(X_J = V).$$

现在，如果令 $\Delta\mu = \infty$，这些特征就只是 V 在 X中不含间隙的出现次数。根据定理 25 和命题 18，该核具有离散质量。 ■

相比之下，标准 k-mer 谱核为 $k(X,\ Y) = \sum_{|V| \le L} u_{\text{la},\,V}(X) u_{\text{la},\,V}(Y)$（例1）。也就是说，它对 K-mer 使用有限和，而非无限和，因此不具备任何灵活性保证。无限 k-mer 谱核具有很强的灵活性保证，同时不会显著牺牲计算效率：我们可以使用局部比对核所用的相同动态规划方法来计算它，因为命题 27 表明它是一种局部比对核（Saigo 等人，2004年）。

10. 嵌入核

本节研究嵌入核，这类核使用欧几里得空间中的序列表示来判断序列相似性。此类核的一个重要动机是避免手工设计相似性度量的需要，转而从数据中学习相似性度量。用于学习表征的数据不必与应用该核的数据相同；例如，一种常见的迁移学习技术是使用大型、未标注的序列数据集来学习序列表示，然后在小型标注数据集上学习从表示到标签的回归（Yang 等人，2018年；Alley 等人，2019年；Biswas 等人，2021年；Rao 等人，2019年；Detlefsen 等人，2022年）。现代流行的无监督表示学习方法包括变分自编码器、循环神经网络和变换器。人们经常发现，这类技术能够生成很好地反映生物学特性的低维表示，即表示相似的序列具有相似的表型。

一个 embedded嵌入核由应用于序列表示的连续核组成。

定义 28 (嵌入核)。设 $F : S \to \mathbb{R}^D$ 是一个嵌入函数，它将序列映射为表示。嵌入核定义为 $k(X,Y) = k_E(F(X)\ F(Y))$，其中 k_E是 $\mathbb{R}^D \times \mathbb{R}^D$上的一个核。

例如，如果我们使用变分自编码器来学习表征，并使用高斯过程来预测标签，那么 F 将由编码器给出，而 k_E 将是高斯过程中的核

。深度核学习也符合嵌入核的形式，其中F 是一个神经网络（Wilson 等，2016年）。如果我们使用神经网络直接从序列回归到标签，那么可以将从序列空间到网络最终隐藏层的映射视为嵌入函数 F，而从隐藏层到输出的映射则近似于核回归；其依据是，在无限宽度极限下，单层神经网络会收敛到高斯过程（Neal，1996年；Williams，1996年）。

10.1 全域性与特征性

我们首先介绍嵌入核满足全域性和特征性的条件。直观而言，如果我们在欧几里得空间中选择一个具有全域性和特征性的核 k_E，那么只要嵌入函数 F 将序列空间 S 中的每个序列映射到 $\mathbb{R}^D$ 中唯一的点，序列空间上的嵌入核也将具有全域性和特征性。

命题 29（通用且具有特征性的嵌入核）。假设 k_E（z、z'）=Ψ（$z - z'$）是一个平移不变核，并且 Ψ 是定义在 $\mathbb{R}^D$ 上的正连续函数，其傅里叶变换严格为正。[5] 那么，当且仅当 F 是单射时，嵌入核 k 才具有特征性和 C_0-通用性。

证明见附录 L。请注意，k_E 的条件确保其在 $\mathbb{R}^D$ 上具有全域性和特征性（Sriperumbudur 等，2010年、2011）。

F 是单射这一条件与某些表示学习方法相冲突，尤其是那些旨在学习稀疏、可解释特征的方法。为了提高可解释性，人们往往希望学习一种能够完全忽略数据点某些特征的表征，例如训练数据中变化很小的特征。然而，当将这种表征用于对留出数据进行预测时，那些被忽略的特征存在变化的序列可能会被映射到嵌入空间中的同一个点。因此，由于嵌入 F 不是单射，下游监督预测器的灵活性将受到限制。附录 M 给出了一个例子。

相比之下，完全“不可解释”的表征并不试图依据生物学意义上的序列相似性来组织嵌入空间，却很容易产生通用且具有特征性的嵌入核。特别是，如果我们使用一个将序列映射为随机表征的嵌入函数，且这些表征服从 $\mathbb{R}^D$上的连续分布，那么任意两个序列具有相同嵌入的概率为零。

命题30（随机嵌入产生通用且具有特征性的核）。假设{F（X）}$_{X\in S}$ 是独立同分布的随机变量，其分布对应的测度相对于勒贝格测度绝对连续。则 F 几乎必然是单射，因此基于 F 的嵌入核几乎必然是通用且具有特征性的。

证明 为 S 中的元素编号，使得 $\{X_1, X_2, \ldots\} = S$。对于任意 $n \in \mathbb{N}$，我们有 $F(\{X_1, \ldots, X_n\})$ 是一个测度为零的集合，因此 $p(F(X_{n+1}) = F(X_i)$ 对某个 $i \leq n) = 0$。于是，$p(F(X_i) = F(X_j)$ 对某个 $i, j) = 0$。 ■

5. 若对所有 ξ 都有$\hat{\Psi}(\xi) > 0$，则 Ψ 的傅里叶变换严格为正，其中$\hat{\Psi}$ 是 Ψ的傅里叶变换。例如，$\Psi(x) = \exp(-x^2)$ 的傅里叶变换严格为正。

该命题的一个推论是，通过对大型神经网络进行随机训练所引入的“无意义”随机性，可能有助于它们生成对下游任务有用的表示。

10.2 离散质量

到目前为止，我们已经描述了嵌入核具有特征性和通用性的条件；我们发现，随机嵌入就足够了。然而，我们在例 3 中看到一个嵌入核，它既通用又具有特征性，却无法可靠地用于优化，因为它不能度量化分布空间。本节将探讨嵌入核何时具有离散质量，从而度量化 $\mathcal{P}(S)$。我们发现，许多标准的表示学习技术很可能会生成不具有离散质量的嵌入核，即使它们具有通用性和特征性也是如此；我们还提出了一种解决办法。

若要使嵌入核具有离散质量，嵌入函数 F 不仅必须将不同的序列映射到 $\mathbb{R}^D$ 中的不同点，这些点还必须充分分散。

命题 31 (具有离散质量的嵌入核**).** 考虑一个具有单射 F 且满足命题 *29* 条件的嵌入核。k_E 度量化 $\mathcal{P}(S)$ 当且仅当 k_E 具有离散质量；而这当且仅当 $F(S)$ 没有聚点时成立。换言之，不存在这样的 $X \in S$，使得 $F(X)$ 属于 $F(S \setminus \{X\})$ 的闭包。

证明见附录 L。在例 3 中，我们看到了一个通用且具有特征性的嵌入核，但它在序列 A 处有一个聚点，因此无法度量 $\mathcal{P}(S)$。

与全域性和特征性不同，要构造具有离散质量的嵌入核，仅使用从 $\mathbb{R}^D$ 上的某个分布中独立同分布抽取的序列表示是不够的。

命题 32 （独立同分布嵌入不具有离散质量）。 假设 $\{F（X）\}_{X\in S}$是独立同分布随机变量。那么，嵌入函数几乎必然存在聚点，因此无法度量 $\mathcal{P}（S）$。

证明 考虑任意 $X_0 \in S$，并注意到几乎必然地，对于任意 X， ϵ， $p(\|F(X) - F(X_0)\| < \epsilon) > 0$。因此，对于所有 ϵ， $p(\|F(X) - F(X_0)\| < \epsilon$对某个 $X) = 1$。所以，以概率 1，存在序列 X_{n_1}， X_{n_2}, ……，它们不是 X_0，但满足当 $k \to \infty$时， $F(X_{n_k}) \to F(X_0)$。
■

直观而言，要构造一个没有聚点的嵌入，就必须设法将序列空间 S 嵌入 $\mathbb{R}^D$，同时避免各个序列的表示彼此过于接近。一种方法是让越长序列的嵌入越分散，而不是从同一分布中采样所有嵌入。

命题 33 (缩放随机嵌入具有离散质量**)**。 考虑一个初始嵌入 F，其中每个 $F(X)$ 对于 $X \in S$，均从球面上的均匀分布中抽取，$\{x \in \mathbb{R}^D \,|\, \|x\| \leq 1\}$。那么，对于 $\epsilon > 0$，使用缩放嵌入 $F(X) = |\mathcal{B}|^{(1+\epsilon)|X|/D}\tilde{F}(X)$的核几乎必然度量化 $\mathcal{P}(S)$ 。

证明见附录 N。直观地说，这种重新缩放为每个序列表示提供了“足够的空间”。如果 $V(B(1))$ 是 $\mathbb{R}^D$ 中单位球的体积，我们可以将 F 视为把集合 $\mathcal{B}^L$ 嵌入半径为 $|\mathcal{B}|^{(1+\epsilon)L/D}$ 的球中。因此，每个序列平均分得的体积为 $V(B(1))(|\mathcal{B}|^{(1+\epsilon)L/D})^D/|\mathcal{B}|^L = V(B(1))|\mathcal{B}|^{L\epsilon}$，它随 L 增大。这恰好提供了足够的空间，从而确保不会发生聚集。

实际上，命题 32 表明，许多表示学习方法不太可能得到能够度量 $\mathcal{P}(S)$ 的核。通常，随着我们嵌入越来越多的序列，它们的表示会继续聚集在嵌入空间中靠近原点的局部区域内，而不是彼此发散。事实上，在表示学习的许多应用中，这种行为是可取的，因为它使表示更易于可视化和解释。理解这种行为的一种方式，是将其视为表示学习中通常采用的先验和正则化器所导致的结果。例如，在潜变量模型中，我们希望整体来看，嵌入的集合 $\{F(X)\}_{X\in S}$ 大致像从潜变量的先验中采样得到的一组样本；由于该先验通常被选为定义在 $\mathbb{R}^D$ 上的连续分布，例如正态分布，命题 32 表明，嵌入核不会具有离散质量。我们在训练数据中只能观测到有限数量的数据点，这一事实也可能造成不利影响：即使训练数据的表示分布得很开，远离训练数据的序列的表示也很可能大致表现得像从先验中采样得到的样本。简言之，典型的表示学习过程不太可能为序列提供足够的空间，从而产生具有离散质量的嵌入核。

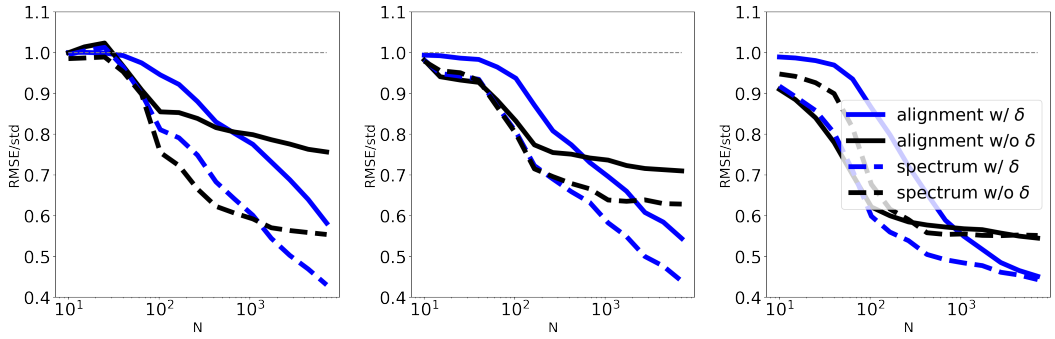
命题 33 提出了一种简单的修正方法：重新缩放嵌入。给定一个嵌入F，其分散程度不会随序列长度增加而提高，可将其修改为$\tilde{F}(X) = |\mathcal{B}|^{(1+\epsilon)|X|/D}F(X)$，其中 $\epsilon > 0$。根据命题 33 的推理，所得的核很可能具有离散质量。在第11节.3 节中，我们展示了这一技巧在实践中的效果。另一种方法是在最初学习嵌入时，使用依赖于长度的先验或正则化。

11. 实证结果

本节考察我们提出的若干具有离散质量的核在真实生物序列数据集上的表现。我们将每个核与一种现有核进行比较；后者依赖于类似的序列相似性概念，但不具有离散质量。我们发现，关于核灵活性的理论保证能够系统性地提升有限数据上的性能。代码见 https://github.com/AlanNawzadAmin/ Kernels-with-guarantees/。

11.1 Predicting transcription factor binding

我们首先考虑一个回归问题，目标是预测某个转录因子是否会与一条 DNA序列结合。理解转录因子结合对于理解和调控基因表达至关重要。这些数据由 DNA序列与转录因子结合强度构成，结合强度通过微阵列实验中的荧光信号强度测量（Barrera等，2016）。我们考察了来自三种不同转录因子的三个独立数据集：同源异型盒蛋白 Hox-D13、Nkx25 和 Esx1。所有 DNA序列的长度均为 8。我们使用核回归拟合数据，并进行评估。



(a) Esx1 (b) Nkx25 (c) Hox-D13

图2：**用于预测转录因子结合的核回归**核回归性能比较：含离散质量与不含离散质量的比对核和 k-mer 谱核（“含δ”与“不含 δ”）。 N 是子抽样数据集的大小。性能以结果变量标准差归一化后的均方根误差衡量。

在训练数据上使用均方根误差进行评估。我们的目标是比较核灵活性；灵活性更高的核应能更好地拟合数据。

我们首先考察比对核，并应用定理 23，将含离散质量的核（$2\mu < \log \sigma$）与不含离散质量的核（$2\mu > \log\ \sigma$）进行比较。我们发现，在所有三个示例数据集的大数据区间内，含离散质量的版本表现更好（图2）。随着数据量增加，不含离散质量的版本最终开始趋于平稳，而含离散质量的版本则稳定地趋近于正确答案，如图2b和图2c所示。

接下来我们考察 k-mer 谱核，并再次比较具有离散质量的核（无限 k-mer 谱核，命题 27）与不具有离散质量的核（例 1 中的有限 k-mer 谱核，含 $L = 3$）。同样，我们发现，在大数据区间，具有离散质量的版本表现更好；而不具有离散质量的版本可能会停滞在较高的误差水平（图 2）。在比对核的情形中，具有离散质量的版本在小数据区间的表现稍差；而在这里，在所考察的数据集大小范围内，具有离散质量的版本表现不逊于、甚至优于不具有离散质量的版本。简言之，确保离散质量能够带来更可靠的回归，尤其是在大数据区间，但并非仅限于此。

11.2 区分免疫受体库

本节讨论一个检验问题，旨在区分不同患者的T细胞受体（TCR）库。T细胞在人类适应性免疫系统中发挥核心作用，它们通过TCR识别外来抗原，进而触发免疫反应。在这里，我们通过比较两位患者TCR序列的分布来比较他们的免疫系统。对于每位患者，我们都有一个由TCR CDR3序列构成的数据集，这些序列长度为10至19个氨基酸（10x Genomics, 2022年）。我们采用

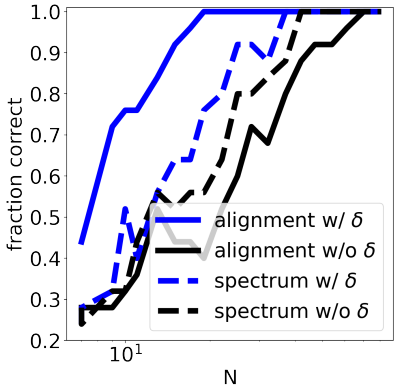


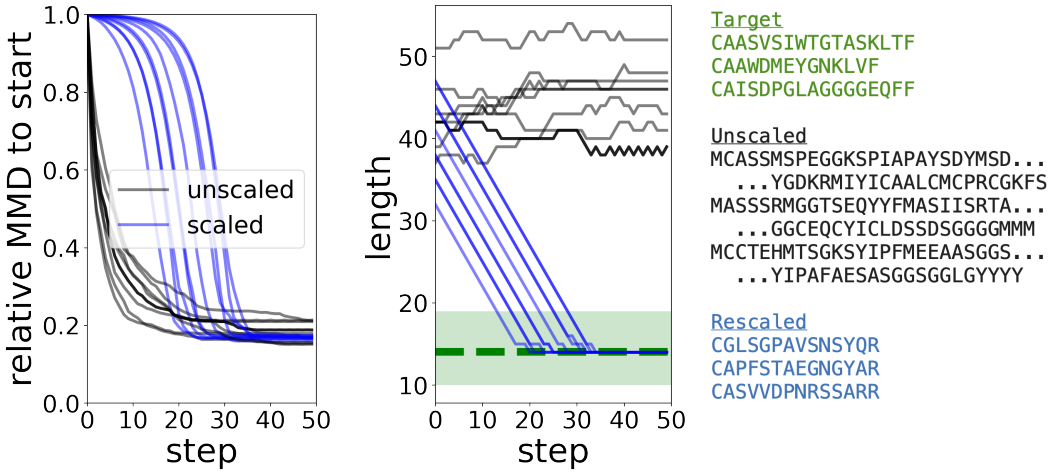
图3：**最大均值差异双样本检验在人类TCR库上的应用**。最大均值差异双样本检验的表现：比较带有和不带有离散质量的比对核与 k-mer 谱核（“带δ”与“不带 δ”）。 N 表示子抽样数据集的大小。表现以独立数据子样本中零假设被正确拒绝的比例衡量。

一种双样本检验：针对最大均值差异（MMD）进行退化U统计量的自助法检验，零假设为两位患者的TCR来自同一分布（Arcones和Gine，1992年）。我们根据检验在90%置信度下正确拒绝零假设的频率来评估性能，并对一个大型数据集进行25次随机子抽样后取平均值。我们的目标是比较核灵活性；灵活性更高的核应当能更可靠地区分两位患者的TCR库。

我们使用第11.1节中的相同四种核。尽管所有核最终都能在数据量足够时区分这两个分布，但具有离散质量的核只需少得多的数据就能可靠地区分它们（图3）。换言之，保证离散质量能够提高双样本检验在真实数据集上的检验效能。

11.3 优化代表性受体

本节讨论一个分布优化问题，目标是找到一个能够代表较大集合的TCR序列。生物学家通常希望了解许多序列的特性，例如浸润患者肿瘤的所有T细胞的TCR，但实验资源有限，因此只能在实验室中合成并测试少量序列。这里，我们旨在找到一个序列 X，使其能够代表一个由大量TCR序列组成的集合 Y_1、……、Y_N。为此，我们优化 δ_X，使其接近经验分布 $p_Y = \frac{1}{N}\sum_{n=1}^{N}\delta_{Y_n}$，优化目标为 $\mathrm{MMD}(\delta_X, p_Y)$。我们使用 $N = 100$ 条人类TCR CDR3序列作为目标集，其长度介于10至17个氨基酸之间（10x Genomics，2022年）。我们将 X 初始化为一个长度超过目标集序列、但氨基酸组成和排列相似的序列（具体而言，从目标集中随机选取一条随机CDR3并将其自身拼接）。随后，在每一步中，通过对单个氨基酸进行替换、插入或删除，优化 $\mathrm{MMD}(\delta_X, p_Y)$，共进行100步。我们的目标是比较各核对度量化 $\mathcal{P}(S)$的能力。由于初始 X 与



(a) (b) (c)

图4：**使用嵌入核优化代表性TCR序列**。分布优化性能，对比未缩放的嵌入核（可能没有离散质量）与缩放的嵌入核（可能具有离散质量）。（a）优化过程中每一步的MMD。此处，MMD根据初始序列 X的MMD进行归一化。使用10个不同的初始序列重复进行优化。（b）优化过程中优化后的序列的长度。目标序列长度的最大值、最小值和平均值以绿色显示。（c）目标集中的序列示例和优化后的序列示例。

目标集的长度；我们特别感兴趣的是，优化 $\mathrm{MMD}(\delta_X, p_Y)$ 是否会得到长度与目标集相似的序列 X。

我们重点研究嵌入核。我们使用预训练嵌入函数 $\tilde{F}: S \to \mathbb{R}^{64}$"UniRep"，该函数由深度循环神经网络学习得到；该网络在一个大型数据集上训练，数据集包含从进化历程中广泛采集的多样化蛋白质序列（Alley 等人，2019年）。我们比较直接使用 F 的嵌入˜核（根据命题 29，这类核很可能是通用且特征性的；但根据命题 32，它们不太可能具有离散质量）与使用缩放嵌入 $F(X) = 20^{(1+\epsilon)|X|/64}\tilde{F}(X)$ 和 $\epsilon = 0.1$ 的嵌入核（根据命题 33，这类核很可能具有离散质量）。对于这两种嵌入函数，我们都使用欧几里得逆多二次核构造嵌入核，即 $k_E(x, y) = (1 + \|x - y\|_2^2)^{-1}$；该核满足命题 29 的条件。对于这两种嵌入核，我们的优化流程在前 50 步内均收敛于最大均值差异（图 4A）。

未缩放核产生的优化后的序列 X 比目标集中的序列长得多，而缩放核可能具有离散质量，其产生的优化后的序列 X 长度与目标集的平均长度相当（图 4B）。对优化后的X进行目视检查也表明，缩放核生成的设计具有明显更强的代表性；例如，缩放核生成的优化后的序列在第一个位置上含有半胱氨酸，与所有目标序列一样，而未缩放核生成的优化后的序列则没有（图 4C）。换言之，对核进行重新缩放以加入离散质量，在实践中提升了其对 $\mathcal{P}(S)$ 进行度量的能力。

12. 讨论

在本文中，我们研究了用于生物序列的核的灵活性。我们的目标是寻找通用、具有特征性并能度量 $\mathcal{P}(S)$ 的核，从而保证它们应用于涉及回归、分布比较和分布优化的机器学习问题时具有可靠性。我们发现，许多现有核无法满足其中一项或全部这些标准；但如果核具有离散质量，则三个标准均可满足。随后，我们对现有核进行了修改，以确保其具有离散质量，从而赋予它们强有力的理论保证，并改善其经验性能。

在欧几里得空间上的应用中，核灵活性问题如今已很少被讨论，而基于核的方法（如高斯过程）的可靠性也常被视为理所当然。这是有充分理由的：软件包中实现、论文和教程中推荐的默认核通常至少是通用且具有特征性的（例如高斯核），而且往往还能度量弱收敛（例如 Mat´ern 核和逆多二次核）（Sriperumbudur 等，2010，2011；Balandat 等人，2020年）。我们致力于为生物序列数据上的核方法提供同样稳固的基础。

我们考察了各种不同的生物学序列相似性概念，但在某些情况下，人们可能还需要其他概念。序列可以首尾相接，如环状肽和环状RNA；也可以组合成大型蛋白质复合物，由许多独立的多肽链构成。序列还可以具有对称性，例如重复序列；也可以发生复杂的突变，例如免疫受体发育过程中的突变。未来的一个研究方向是构造适用于这些或其他生物学现

象的具有离散质量的核。我们为证明核具有离散质量而开发的工具（第6节）提供了一个起点。

另一个方向是开发能够从数据中自动学习复杂核的方法，同时保留离散质量性质。我们在第7至10节中的证明展示了如何从一个具有离散质量的简单核出发，应用一系列保持离散质量的变换（第6.2节），开发出一个具有离散质量的更复杂核。学习新核的一种可能方法是优化这一系列变换。

总的来说，我们的工作为有关生物序列机器学习理论基础的规模虽小但日益增长的文献作出了贡献。针对生物序列的机器学习方法对人类健康和环境的重要性日益提高，而且通常应用于信任与可靠性至关重要的问题。然而，人们有时会错误地将其视为“通用”方法的特例，认为这些方法原本是为其他类型的数据发明的。相反，我们的研究结果鼓励人们进一步探究生物序列分析的专门理论。

A. 具有有限特征的核不具有特征性

命题 34。（命题 7 的证明）考虑一个使用有限特征向量定义的核 $k(X,Y) = (\phi(X) \mid \phi(Y))$ $\phi : S \to \mathbb{R}^d$。如果 S 是无限的，则该核不具有特征性。

证明 首先，我们将修改 ϕ，使得特征几乎不会落在 $\mathbb{R}^d$ 的仿射子空间中，也就是说，避免存在 $v \in \mathbb{R}^d \setminus \{0\}$ 和 $\beta \in \mathbb{R}$，使得对于除有限多个 $X \in S$ 之外的所有情况，$(v|\phi(X)) = \beta$。如果 ϕ 确实具有这样的性质，那么存在一个酉变换 U，使得对于除有限多个 $X \in S$ 之外的所有情况，第一坐标 $(U\phi(X))_1 = \|v\|\beta$。定义到 $\mathbb{R}^{d-1}$ 的嵌入：$\tilde{\phi}(X) = (U\phi(X))_{2:}$，使得$k(X,Y) = (U\phi(X)|U\phi(Y)) = \|v\|^2\beta^2 + (\tilde{\phi}(X)|\tilde{\phi}(Y))$ 对所有 $X, Y \in \tilde{S}$ 都成立，其中 $\tilde{S}$ 是 S 的一个无限有限子集。令 $k(X,Y) = (\phi(X)|\phi(Y))$，注意，对于任意分布 μ, ν，

$$\begin{aligned}\mathrm{MMD}_k(\mu,\nu) =& E_{X\sim\mu, Y\sim\nu}\left[k(X,X) + k(Y,Y) - 2k(X,Y)\right] \\ =& E_{X\sim\mu, Y\sim\nu}\left[\tilde{k}(X,X) + \tilde{k}(Y,Y) - 2\tilde{k}(X,Y)\right] = \mathrm{MMD}_{\tilde{k}}(\mu,\nu)\end{aligned}$$

~ ~ ~ ~因此，k 作为 S 上的核具有特征性，当且仅当 k 具有特征性。将 k 重新定义为 k，将 S 重新定义为 S，将~重新定义为 ϕ，并将 ϕ 重新定义为 $v \in \mathbb{R}^d$，然后根据需要重复上述构造任意多次，我们可以假设，对于任意 $\beta \in \mathbb{R}$、$X \in S$，都存在无穷多个$v|\phi$使得（X（$\neq \beta$））ϕ，也就是说，S（$\mathbb{R}^d$）不属于的任何仿射子空间。

现在我们将通过归纳法，在 $d+1$ 中找出 ϕ（S）个仿射无关点，以构造一个内部非空的凸集合。设 X_0、$X_1 \in S$ 满足 ϕ（X_1）$\neq \phi$（X_0）。假设我们已选取 X_0、.....、X_n，使得 $\{\phi$（X_i）$-\phi$（X_{i-1}）$\}_{i=1}^n$ 对于 $n < d$构成一个线性无关集合。记$e_i = \phi$（X_i）为 $-\phi$（X_{i-1}）。由于 $\{e_i\}_{i=1}^n$ 是一个由 $n < d$ 向量组成的集合，我们可以选取 $v \in \mathbb{R}^d$，使得对所有 $i \le n$，均有（$v|e_i$）$= 0$。根据上一段中的假设，我们可以选取一个 X_{n+1}，使得（$v|\phi$（X_{n+1}）$-\phi$（X_n））$\neq 0$，这尤其意味着 $\{\phi$（X_i）$-\phi$（X_{i-1}）$\}_{i=1}^{n+1}$线性无关。因此，我们已选取 $\{X_i\}_{i=0}^d$，使得 $\{e_i\}_{i=1}^d$ 线性无关。

设 C 是 $\{\phi(X_i)\}_{i=0}^d$ 的凸包。设 $\mu_0, \ldots, \mu_d$ 是一组总和为 1 的严格正数，且 $w = \sum_i \mu_i \phi(X_i) \in C$。注意，对于任意满足 $|c_i| < \min_i \mu_i$ 的 $c_1, \ldots, c_d \in \mathbb{R}^d$，我们有

$$w + \sum_i c_i e_i = (\mu_0 - c_0)\phi(X_0) + (\mu_1 + c_0 - c_1)\phi(X_1) \cdots + (\mu_d + c_d)\phi(X_d) \in C.$$

特别地，由于 $\{e_i\}_i$ 是 $\mathbb{R}^d$的基，w 位于 C的内部。现在令 $Z \in S$ 对所有 i具有$Z \neq X_i$。对于足够小的 $\epsilon > 0$，$\epsilon\phi(Z) + (1-\epsilon)w \in C$，特别地，存在一个和为 1 的正数集合，$\nu_0, \ldots, \nu_d$，使得

$$\epsilon\phi(Z) + \sum_i (1-\epsilon)\mu_i \phi(X_i) = \sum_i \nu_i \phi(X_i).$$

因此，k不是特征性的：分布 $\epsilon\delta_Z + \sum_i (1-\epsilon)\mu_i \delta_{X_i}$和 $\sum_i \nu_i \delta_{X_i}$ 的核嵌入相同。

■

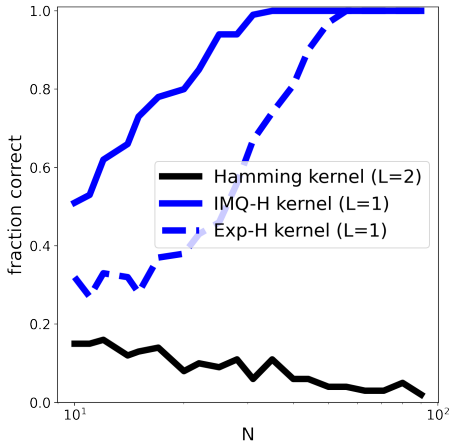


图 5：将最大均值差异双样本检验应用于模拟数据。我们比较滞后 $L = 2$ 的汉明核、指数汉明核（Exp-H）和逆多二次汉明核（IMQ-H）。只有后两者具有离散质量。

B. 滞后 L 的汉明核无法区分分布

非特征性核能够区分的分布有限。这里，我们使用滞后为 L 的汉明核（第 3 节）通过模拟对此进行说明。

我们考虑两个定义在长度为 10 的 DNA序列上的分布：第一个是均匀分布；第二个在前 5 个字母上服从均匀分布，而后 5 个字母与前 5 个相同。我们使用不同的核，评估最大均值差异双样本检验区分这两个分布的能力。第一个核是滞后 $L = 2$ 的汉明核，它不具有特征性。其他核为指数汉明核（滞后为 $L = 1$）和逆多二次汉明核（滞后为 $L = 1$），它们具有离散质量，因此具有特征性（第7节）。对于从这两个分布中分别抽取的、数量逐渐增加的数据，我们使用每种核进行最大均值差异双样本检验；该检验针对退化U统计量使用自助法，零假设为两个分布相同（Arcones和Gine，1992年）。我们将拒绝零假设的置信阈值设为 95%。一个好的检验应当只需少量数据，就能可靠地拒绝零假设。

图 5 展示了在 100 次独立数据抽样中，双样本检验正确拒绝原假设的比例（即检验的功效）。可以看到，无论数据量有多大，滞后为 L 的汉明核都无法可靠地区分这两个分布；其渐近功效为零。然而，数据量足够大时，指数核和 IMQ-H核能够可靠地区分这两个分布。简言之，对标准汉明核稍作改动，为其赋予离散质量，就能保证它不会失效。

C. 能够度量化分布空间但缺少离散质量的通用核

这里我们将证明，要构造一个通用核或能够度量化 $\mathcal{P}(S)$ 的核，并非一定需要离散质量。回顾例 3，其中该核无法度量化 $\mathcal{P}(S)$，因为 k_A 可以在 $\mathcal{H}_k$ 中由 $k_{n\times A}$ 很好地逼近。为了得到一个能够度量化 $\mathcal{P}(S)$ 但不含离散质量的通用核，我们将在整数$\{0, 1, -1, 2, -1, \dots\}$上构造一个核，使得 k_0可以由 $\int d\nu_n(X)k_X$逼近 $\nu_n \in \mathcal{P}(S)$，但前提是 ν_n 的全变差随 n 增大而趋于无穷。

令 $\mathcal{H}_k$ 为一个希尔伯特空间，其标准正交基由整数$\{e_0, e_1, e_{-1}, \dots\}$索引。定义 $k_0 = e_0$，对于任意正的 n，$k_n = e_n$以及 $k_{-n} = \sqrt{1-\epsilon_n^2-\epsilon_n^4}e_n + \epsilon_n^2 e_{-n} + \epsilon_n e_0$，其中 ϵ_n是一个递减的正 ϵ_n序列，且满足 $\epsilon_n \to 0$。显然，k是定义在整数上的严格正定核，并且对所有整数z，$k(z,z) = 1$。

$\mathcal{H}_k$ 不能包含 δ_0。为说明这一点，假设 $\delta_0 \in \mathcal{H}_k$。那么

$$\begin{aligned}
0 =& \frac{1}{\epsilon_n}\left(\delta_0(-n) - \delta_0(n)\right) \\
=& \left(\delta_0 \middle| \frac{1}{\epsilon_n}\left(k_{-n} - k_n\right)\right)_k \\
=& \left(\delta_0 \middle| \frac{\sqrt{1-\epsilon_n^2-\epsilon_n^4}-1}{\epsilon_n} e_n + \epsilon_n e_{-n} + e_0\right)_k \\
\to& \left(\delta_0 | e_0\right)_k = \delta_0(0) = 1,
\end{aligned}$$

矛盾。

然而，k 是 C_0-通用的。为说明这一点，假设对于定义在 S上的某些有限z（可能带符号）测度 $\epsilon_n^2\pi_{-n} \quad 0 = \sum_z \pi_z k_z|e_n\rangle_k = \pi_n + \pi_{-n}\sqrt{1-\epsilon_n^2-\epsilon_n^4} \quad \pi_n = 0$，有 $\sum_z \nu(z)k_z = \sum_z \mu(z)k_z$。记作 。注意，$0 = (\sum_z \pi(z)k_z|e_{-n})(z)$，且 $(z)(z)(z)$。因此，对所有$n \neq 0$，(z)。最后，$0 = (\sum_z \pi(z)k_z|e_0)_k = \pi(0)$，所以 $\pi = 0$，即 $\mu = \nu$。根据 Sriperumbudur 等人（2011）的命题 2，k 是通用的。

k也将 $\mathcal{P}(S)$度量化。为此，设对于某个 μ，ν_1，ν_2，$\dots \in \mathcal{P}(S)$，我们有$\sum_z \nu_m(z)k_z \to \sum_z \mu(z)k_z$。记 $\pi_m = \nu_m - \mu$。如上所述，$(\sum_z \pi_m(z)k_z|e_{-n})_k = \epsilon_n^2\pi_m(-n)$以及$(\sum_z \pi_m(z)k_z|e_n)_k = \pi_m(n) + \pi_m(-n)\sqrt{1-\epsilon_n^2-\epsilon_n^4}$。对于每个 n，这两个方程的左侧均趋于 0，当 $m \to \infty$ 时，因此，$\pi_m(n) \to 0$ 对于所有 $n \neq 0$。最后注意

$$\left(\sum_z \pi_m(z)k_z \middle| e_0\right)_k = \pi_m(0) + \sum_{n=1}^{\infty}\pi_m(-n)\epsilon_n.$$

左侧趋于 0，并且由于 $\sum_z |\pi_m(z)| \leq 2$ 以及 $\epsilon_n \to 0$，右侧的和也趋于 0。因此 $\pi_m(0) \to 0$，从而 $\nu_n \to \mu$。

有趣的是，尽管 $\delta_0 \notin \mathcal{H}_k$，$k$ 仍能在无穷范数下任意精确地逼近 δ_0。首先，对每个 n，选取一个 $f_n \in$ 张成空间$\{k_0, k_n, k_{-n}\}$，使得 $(f_n|k_0) = (f_n|k_n) = 0$ 且 $(f_n|k_{-n}) = 1$。然后，$g_n = k_0 - \sum_{n=1}^N \epsilon_n f_n$在 $\{0, N+1, -(N+1)\ N+2,\ -(N+2)\ \dots\}$上为 k_0，在 $\{1, -1, \dots, N, -N\}$上为 0。由于 k_0在无穷远处趋于零，g_n在无穷范数下收敛于 δ_0

当 $N \to \infty$ 时， δ_0 在无穷范数下收敛。然而，尽管 $(g_n)_n$ 在无穷范数下逼近 δ_0 ，它并未在 $\mathcal{H}_k$中逼近 δ_0 ： $\|g_n\|_k$ 无界增长。[6]

D. 经处理的核的再生核希尔伯特空间

此处给出第6节中各命题的证明

命题 35. $f \mapsto Af$ 是从 $\mathcal{H}_k$ 到 $\mathcal{H}_{k^A}$的酉同构。

证明 在$\{k_Y\}_{Y\in S}$和 T的张成空间上定义一个线性变换，将 k_x映射到$A(X)^{-1}k_x^A$。T定义良好，因为若对于某个X_n有 $\sum_{n=1}^N \alpha_n k_{X_n} = 0$，且$_n$、$_n$分别为$(\alpha_n)$、$(\sum_{n=1}^N A)$，则对于所有 Z，都有$X_n{}^{-1}\alpha_n k_{X_n}^A|k_Z^A)$ $_{k^A} = A(Z)\sum_{n=1}^N \alpha_n k(X_n, Z) = 0$。 T也是酉的，因为$(Tk_X|Tk_Y)_{k^A} = k(X,Y) = (k_X|k_Y)_k$。因此， T可以扩展到整个 $\mathcal{H}_{k^A}$。

注释 $(Tk_X|k_Y^A)_{k^A} = k(X,Y)A(Y) = (k_X|k_Y)_k A(Y)$ 对于所有 $X, Y \in S$。由线性性和连续性，对于所有 $f \in \mathcal{H}_k$, $Tf(Y) = (Tf|k_Y^A)_{k^A} = (f|k_Y)_k A(Y) = f(Y)A(Y)$。

我们可以看到， T 是满射，因为它具有逆矩阵 $f \mapsto A^{-1}f$。 ■

命题 36. 设 k_1, k_2 是 S上的核。定义

$$k_1 \otimes k_2((X_1, X_2), (Y_1, Y_2)) = k_1(X_1, Y_1)k_2(X_2, Y_2)$$

对于(X_1, X_2)， $(Y_1, Y_2) \in S^2$。那么$\mathcal{H}_{k_1\otimes k_2} = \mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2}$。特别地，如果 $f_1 \in \mathcal{H}_{k_1}, f_2 \in \mathcal{H}_{k_2}$, 则存在一个 $f_1 \otimes f_2 \in \mathcal{H}_{k_1\otimes k_2}$ ，使得对于$(X_1, X_2) \in S^2$,

$$f_1 \otimes f_2(X_1, X_2) = f_1(X_1)f_2(X_2). \tag{9}$$

证明 对于 $f_1 = \sum_{n=1}^{N_1} \alpha_{1,n} k_{1,X_{1,n}} \in \mathcal{H}_1, f_2 = \sum_{n=1}^{N_2} \alpha_{2,n} k_{2,X_{2,n}} \in \mathcal{H}_2$, 定义

$$f_1 \otimes f_2 = \sum_{n=1}^{N_1}\sum_{m=1}^{N_2} \alpha_{1,n}\alpha_{2,m} (k_1 \otimes k_2)_{(X_{1,n}, X_{2,m})} \in \mathcal{H}_{k_1\otimes k_2}. \tag{10}$$

显然，若 $f_1' \in \text{span}(k_{1,X})_{X\in S}, f_2' \in \text{span}(k_{2,X})_{X\in S}$, 则 $(f_1 \otimes f_2|f_1' \otimes f_2')_{k_1\otimes k_2} = (f_1|f_1')_{k_1}(f_2|f_2')_{k_2}$。因此，我们可以将双线性函数 $\otimes$连续延拓到所有$f_1 \in \mathcal{H}_{k_1}, f_2 \in \mathcal{H}_{k_2}$。 $\mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2}$ 由此嵌入 $\mathcal{H}_{k_1\otimes k_2}$, 并且实际上作为$\{k_{1,X_1} \otimes k_{2,X_2}\}_{(X_1,X_2)} = \{(k_1 \otimes k_2)_{(X_1,X_2)}\}_{(X_1,X_2)}$ 具有稠密的张成空间，从而实现满射嵌入。最后，等式 9 可由 $k_{1,X_1} \otimes k_{2,X_2} = (k_1 \otimes k_2)_{(X_1,X_2)}$、等式 10 以及连续性推出。 ■

命题 37。 设 k 是 S 和 $S' \subseteq S$上的核。将 $k|_{S'}$ 定义为 k在 $S' \times S'$上的限制。则 $\mathcal{H}_{k|_{S'}}$ 是 $\mathcal{H}_k$中张成空间$\{k_Y\}_{Y\in S'}$ 的闭包。

6. 要看出为什么 $\|g_n\|_k$ 必定发散，请注意，如果情况并非如此，那么根据巴拿赫–阿劳格鲁定理，一个子序列 $(g_{n_k})_k$必定“弱收敛”到一个 $g \in \mathcal{H}_k$，也就是说，对于所有 z， $g_{n_k}(z) = (g_{n_k}|k_z)_k \to (g|k_z)_k = g(z)$。因此， $\delta_0 = g \in \mathcal{H}_k$，矛盾。

证明令 T：张成空间$((k|_{S'})_X)_{X\in S'} \to \mathcal{H}_k$ 是一个线性映射，它将$(k|_{S'})_X$ 映射到 k_X，对于 $X \in S'$。 T是良定义的，因为如果 $f = \sum_{n=1}^N \alpha_n (k|_{S'})_{X_n} = 0$ 对于某个 $(X_n)_n$、$(\alpha_n)_n$，那么

$$\|Tf\|_k = \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m k(X_n, X_m) = \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m k|_{S'}(X_n, X_m) = \|f\|_{k|_{S'}} = 0.$$

T 显然也是酉的，因此可以扩展为作用在整个 $\mathcal{H}_{k|_{S'}}$上的酉映射。显然， T 的像是 $\mathcal{H}_k$ 中 $\mathrm{span}\{k_Y\}_{Y\in S'}$ 的闭包。 ■

E. 比对核中的离散质量

本节证明，在某些参数设置下，比对核具有离散质量。

我们首先建立具有非独热编码的序列的一些性质。特别是，我们表明，在对字母表重新参数化时，核的卷积仍以相同方式运作。注意，对于 V， $W \in \cup_{L=0}^\infty \mathbb{R}^{L\times\mathcal{B}}$，我们用 $V_{(l}$, $V_{(l:l')}$ 索引第一维上的位置，并用 $V+W$ 表示沿第一维的拼接，这与 $S = \cup_{L=1}^\infty \mathcal{B}^L$ 中常规序列的定义类似。

命题 38。 设 k_1、k_2 是定义在 S 和 $V \in \mathbb{R}^{L\times\mathcal{B}}$、$V' \in \mathbb{R}^{L'\times\mathcal{B}}$ 上的核。

$$k_1 \star k_2(V, V') = \sum_{l_1+l_2=L, l_1'+l_2'=L'} k_1\left(V_{(:l_1)}, V'_{(:l_1')}\right) k_2\left(V_{(l_1:)}, V'_{(l_1':)}\right).$$

证明

$$\begin{aligned}
&k_1 \star k_2(V, V') \\
&= \sum_{|X|=L} \sum_{|Y|=L'} \left(\prod_{l=0}^{L-1} V_{l,X_{(l)}}\right) \left(\prod_{l=0}^{L'-1} V_{l,Y_{(l)}}\right) \sum_{l_1,l_1'} k_1(X_{(:l_1)}, Y_{(:l_1')}) k_2(X_{(l_1:)}, Y_{(l_1':)}) \\
&= \sum_{l_1,l_1'} \sum_{|X|=L} \sum_{|Y|=L'} \left(\left(\prod_{l=0}^{l_1-1} V_{l,X_{(l)}}\right) \left(\prod_{l=0}^{l_1'-1} V_{l,Y_{(l)}}\right) k_1(X_{(:l_1)}, Y_{(:l_1')})\right) \\
&\qquad \times \left(\left(\prod_{l=l_1}^{L-1} V_{l,X_{(l)}}\right) \left(\prod_{l=l_1'}^{L'-1} V_{l,Y_{(l)}}\right) k_2(X_{(l_1:)}, Y_{(l_1':)})\right) \qquad (11) \\
&= \sum_{l_1,l_1'} \left(\sum_{|X|=l_1} \sum_{|Y|=l_1'} \left(\prod_{l=0}^{l_1-1} V_{l,X_{(l)}}\right) \left(\prod_{l=0}^{l_1'-1} V_{l,Y_{(l)}}\right) k_1(X, Y)\right) \\
&\qquad \times \left(\sum_{|X|=L-l_1} \sum_{|Y|=L'-l_1'} \left(\prod_{l=0}^{L-l_1-1} V_{l+l_1,X_{(l)}}\right) \left(\prod_{l=0}^{L'-l_1'-1} V_{l+l_1',Y_{(l)}}\right) k_2(X, Y)\right) \\
&= \sum_{l_1,l_1'} k_1\left(V_{(:l_1)}, V'_{(:l_1')}\right) k_2\left(V_{(l_1:)}, V'_{(l_1':)}\right).
\end{aligned}$$

■

现在我们证明定理 23。

定理 39. （定理 *23* 的证明）定义 K 为矩阵，其元素为 $K_{b',\ b} = k_s(b, b')$ ，其中$b, b' \in \mathcal{B}$，且 $\sigma = \mathbf{1}_{\mathcal{B}}^T K^{-1} \mathbf{1}_{\mathcal{B}}$。如果 $\Delta\mu > 0$，那么 k具有离散质量，当且仅当$2\mu \geq$对数σ。如果 $\Delta\mu = 0$，那么 k具有离散质量，当且仅当 $2\mu >$对数 σ。

证明 证明分为以下几个步骤。首先，我们对 $\mathcal{B}$ 重新参数化，使得 $\mathcal{H}_k$ 分解为一组正交超平面 W_V。其次，我们证明这些超平面中的每一个都与比对核的再生核希尔伯特空间的张量积同构，且$|\mathcal{B}| = 1$。最后，根据定理 45 可得结果；该定理给出了比对核具有 $|\mathcal{B}| = 1$ 时存在离散质量的条件。

*第 1 部分：*我们从构造重新参数化字母表开始。令 $\tilde{U} = K^{-1}\mathbf{1}_{\mathcal{B}}$ 为$\mathcal{B}$每个位置都为1的向量，并令 $U = \tilde{U}/k_s(\tilde{U}, \tilde{U})$，因此，$U^T \mathbf{1}_{\mathcal{B}} = \tilde{U}^T K \tilde{U}/\tilde{U}^T K \tilde{U} = 1$ 和 $k_s(U, U) = (\tilde{U}^T K \tilde{U})^{-1} = k_s(\tilde{U}, \tilde{U})^{-1} = (\mathbf{1}_{\mathcal{B}}^T K^{-1} \mathbf{1}_{\mathcal{B}})^{-1} = \sigma^{-1}$。令 $B_1, \ldots, B_{|\mathcal{B}|-1} \in \mathbb{R}^{\mathcal{B}}$ 满足：$\{U, B_1, \ldots, B_{|\mathcal{B}|-1}\}$ 是向量空间 $\mathbb{R}^{\mathcal{B}}$ 的一个正交基，其点积为 $(v|x) = v^T K w$，且对所有 i，$(B_i|B_i) = 1$ 。因此，对所有 i, j，$k_s(B_i, B_j) = \delta_i(j)$；并且对所有 i，$k_s(U, B_i) = 0$。现在令 $\mathcal{B} = \{U, B_1, \ldots, B_{|\mathcal{B}|-1}\}$。~我们将使用 $\mathcal{B}$ 作为重新参数化字母表，并注意到根据命题 19，比对核具有离散质量，当且仅当将 $\mathcal{B}$ 替换为 $\mathcal{B}$ 时，比对核仍具有离散质量。

现在，对于任意数 L 和 $V \in \tilde{\mathcal{B}}^L$，将插入核 k_I 应用于 V，得到：

$$k_{I,V} = \sum_{|X|=L} \left(\prod_{l=0}^{L-1} V_{l,X_{(l)}} \right) e^{-\Delta\mu - L\mu} g = e^{-\Delta\mu - L\mu} \prod_{l=0}^{L-1} \left(\sum_{b \in \mathcal{B}} V_{l,b} \right) g.$$

注意，对于任意 $B \in \{B_1, \ldots, B_{|\mathcal{B}|-1}\} \subset \tilde{\mathcal{B}}$，我们有 $\sum_{b \in \mathcal{B}} B_b = B^T \mathbf{1}_{\mathcal{B}} = B^T K \tilde{U} = k_s(B, U)/k_s(U, U) = \delta_U(B)$。因此，$k_{I,\ V} = 0$如果 V 中的任意字母属于 $\mathcal{B} \setminus \{U\}$。现在设 $V \in \tilde{\mathcal{B}}^L$, $V' \in \tilde{\mathcal{B}}^{L'}$。根据命题 38，$k(V, V')$ 由下式给出：

$$\sum_{l \in \mathbb{N}, V^{(1)} + \cdots + V^{(2l+1)} = V, V'^{(1)} + \cdots + V'^{(2l+1)} = V'} k_I(V^{(1)}, V'^{(1)}) \prod_{i=1}^{l} k_s(V^{(2i)}, V'^{(2i)}) k_I(V^{(2i+1)}, V'^{(2i+1)}).$$

只有当 $V^{(m)}$ 和 $V'^{(m)}$ 仅由字母 U组成（对所有奇数 m 均如此），且 $V_{(m)} = V'^{(m)}$和 $|V^{(m)}| = 1$ （对所有偶数 m 均如此）时，求和中的每一项才非零。特别地，如果$V \in (\tilde{\mathcal{B}} \setminus \{U\})^L$，满足 $k(V, V') \neq 0$ 的序列 V' 必须具有如下形式：

$$(t_0 \times U) + V_{(0)} + (t_1 \times U) + V_{(1)} + \cdots + V_{(L-1)} + (t_L \times U)$$

对于某些整数 t_0、.....、t_L。将此序列称为 $V[t]$。令 $W_V \subset S$表示所有此类序列组成的集合，即 $W_V = \{V[t] \mid t_0, \ldots, t_N \in \mathbb{N}\}$。我们看到，$W_V$与任意满足 $V' \notin W_V$ 的 $W_{V'}$正交。因此，每个 W_V 都定义了一个正交超平面。此外，所有 $\{W_V\}_V$ 的并集~是字母取自 $\mathcal{B}$的所有序列组成的集合。因此，命题 16 表明，如果比对核限制在每个 W_V 上时具有离散质量，那么它就具有离散质量。

第二部分：本节将说明，要使比对核在 W_V 上具有离散质量，只需一个更简单的比对核（其字母表大小仅为 1）具有离散质量。设 $V[t]$、$V[t'] \in W_V$，其中 $V \in (\tilde{\mathcal{B}} \setminus \{U\})^L$。令 $V[t_{1:}]$ 表示去掉最前面的 t_0+1 个字母后的 $V[t]$。由于具有正权重的比对只有一种：对所有 i，将 $V[t]$ 中第 i 组 U 与 $V[t']$ 中第 i 组 U 进行比对，因此我们可以将比对核分解为多个比对核的乘积，分别比较由 U 构成的各个字符串：

$$
\begin{aligned}
k(V[t],V[t']) = & \sum_{\substack{l\in\mathbb{N} \\ Y^{(1)}+\cdots+Y^{(2l+1)}=t_0\times U \\ Z^{(1)}+\cdots+Z^{(2l+1)}=t'_0\times U}} \sum_{\substack{l'\in\mathbb{N} \\ \tilde{Y}^{(1)}+\cdots+\tilde{Y}^{(2l'+1)}=V[t_{1:}] \\ \tilde{Z}^{(1)}+\cdots+\tilde{Z}^{(2l'+1)}=V[t'_{1:}]}} \Bigg(\\
& \times \left(k_I(Y^{(1)},Z^{(1)}) \prod_{i=1}^{l} k_s(Y^{(2i)},Z^{(2i)}) k_I(Y(2i+1),Z^{(2i+1)}) \right) \\
& \times k_s(V_{(0)},V_{(0)}) \\
& \times \left(k_I(\tilde{Y}^{(1)},\tilde{Z}^{(1)}) \prod_{i=1}^{l'} k_s(\tilde{Y}^{(2i)},\tilde{Z}^{(2i)}) k_I(\tilde{Y}^{(2i+1)},\tilde{Z}^{(2i+1)}) \right) \Bigg) \\
= & k(t_0\times U, t'_0\times U) k(V[t_{1:}],V[t'_{1:}]) \\
= & \dots \\
= & \prod_{l=0}^{L} k(t_l\times U, t'_l\times U).
\end{aligned}
\tag{12}
$$

现在，当 $\mathcal{B}=\{A\}$ 时，将 k_1 称为比对核，其参数为 $k_s(A,\ A)=\sigma^{-1}$。由于$k_s(U,\ U)=\sigma^{-1}$，我们有 $k_1(l\times A,\ l'\times A)=k(l\times U,\ l'\times U)$。因此，$k$ 在 W_V 上的作用等同于 k_1 的$L+1$次张量积。因此，根据命题 17，如果 k_1 具有离散质量，那么对于每个 V，k 在 W_V 上也具有离散质量，所以 k 在 S 上具有离散质量；另一方面，如果 k_1 不具有离散质量，那么 k 在 $W_\emptyset$ 上也不具有离散质量，所以 k 在 S 上不具有离散质量。

最后，在定理 45 中我们将看到，如果 $|\mathcal{B}|=1$，那么当 $\Delta\mu>0$ 时，k_1 具有离散质量，当且仅当 $2\mu+\log k(A,\ A)\geq 0$。因此，如果 $\Delta\mu>0$，我们发现 k_1，因而 k 具有离散质量，当且仅当 $2\mu\geq\log\sigma$。对于 $\Delta\mu=0$，结论也类似。 ■

注意，这个证明也适用于对插入不同字母施加不同惩罚的比对核。更确切地说，我们可以通过对函数 $t\in\mathbb{R}^{\mathcal{B}}$使用 $\tilde{g}(X)=\left(\prod_{l=0}^{|X|-1} t(X_{(l)})\right) g(X)$，修改 k_I，使其针对不同字母而有所不同。上文我们选择了$t=\mathbf{1}_{\mathcal{B}}$。对于一般的 t，定义 $U=K^{-1}t/t^TK^{-1}t$，并以相同的方式定义 $\tilde{\mathcal{B}}=\{U,B_1,\dots,B_{|\mathcal{B}|-1}\}^{\sim}$。同样，对于任意 $B\in\mathcal{B}$，$\sum_b t(b)B_b=k_s(B,U)/k_s(U,U)=\delta_U(B)$，因此，对于任何含有与 U 正交字母的序列 V，$k_{I,V}=0$。令 $\sigma=t^TK^{-1}t$，我们可以重复定理 23 的证明，得出相同的结论。

F. Int 解读比对核中离散质量的条件 s

在定理 23 中，我们看到，比对核只有在满足某些条件时才具有离散质量。这些条件的一个有趣之处在于，它们涉及核的插入参数（通过 μ）与替换参数（通过 σ）之间的相互作用。人们通常认为这些参数截然不同；毕竟，从生物学角度看，替换突变和插入突变可能源自截然不同的分子过程。在这里，我们将帮助读者直观理解这些参数如何相互作用并决定离散质量。我们聚焦于 $\Delta\mu = 0$ 这一特殊情形。

核心思路是将比对核重写为对某种特定逐位置比较核在每一对成对比对上的求和；这样一来，比对核具有离散质量的条件，恰好与相应逐位置核具有离散质量的条件相同。具体而言，回顾公式 7 将比对核分解为对成对比对的求和。例如，考虑我们在第 8.1 节中看到的 X =ATGC 与 Y =ACC 之间的比对。我们可以将公式 7 中相应的项改写为

$$k'_s(\mathtt{A},\mathtt{A})k'_s(\mathtt{T},\mathtt{C}), k'_s(\mathtt{G},\mathtt{-})k'_s(\mathtt{C},\mathtt{C})k'_s(\mathtt{-},\mathtt{-})k'_s(\mathtt{-},\mathtt{-})\dots$$

其中，我们将扩展字母比较核定义为 k'_s as $k'_s(b,\ b') = k_s(b,\ b')$，适用于 $b,\ b' \in \mathcal{B}$；$k'_s($ $b,$ -$) = e^{-\mu}$ 适用于 $b \in \mathcal{B}$；以及 $k'_s($-, -$) = 1$。这符合逐位置比较核（定义 20）的形式，该核作用于序列 ATGC--... 和 AC-C--...，其字母核为 k'_s，且空位符号同时也是终止符号。一般而言，构成比对核之和的每一项（公式 7）都可以用这种方式表示，并且使用相同的字母核。

应用定理 21，逐位置核具有离散质量，当且仅当k'_s 是严格正定的。[7] 这又取决于替换参数与插入参数之间的关系，因为两者都会影响 k'_s。令 $V \in \mathbb{R}^{\mathcal{B}\cup\{\text{-}\}}$ 使得V- $= 1$。称 $\tilde{V} = (V_b)_{b\in\mathcal{B}}$ 为 V在 $\mathcal{B}$中的分量。定义 $\tilde{U} = K^{-1}\mathbf{1}_{\mathcal{B}}$ 使得 $\tilde{U}K\tilde{U} = \sigma$。现在，

$$k'_s(V,V) = \tilde{V}^T K\tilde{V} + 2e^{-\mu}\tilde{V}^T\mathbf{1}_{\mathcal{B}} + 1 = \tilde{V}^T K\tilde{V} + 2e^{-\mu}\tilde{V}^T K\tilde{U} + 1.$$

现在将 $\tilde{V}$分解为 $\tilde{V} = \alpha\tilde{U} + \tilde{V}^\perp$，其中 $\tilde{V}^\perp$ 的列在内积（$v|w$） $= v^T Kw$下与$\tilde{U}$ 的列正交。于是，

$$k'_s(V,V) = \tilde{V}^{\perp T}K\tilde{V}^\perp + \alpha^2\tilde{U}K\tilde{U} + 2\alpha e^{-\mu}\tilde{U}K\tilde{U} + 1 = \tilde{V}^{\perp T}K\tilde{V}^\perp + \sigma\left(\alpha^2 + 2e^{-\mu}\alpha\right) + 1.$$

此表达式在 $\alpha = -e^{-\mu},\ \tilde{V}^\perp = 0$处取得最小值，此时

$$k'_s(V,V) = -\sigma e^{-2\mu} + 1.$$

因此，k'_s 是严格正定的，当且仅当 $e^{-2\mu}\sigma < 1$，即 $2\mu > \log\sigma$。所以，比对核具有离散质量，当且仅当逐位置比较核具有离散质量。

归根结底，比对核的灵活性取决于我们用于比较对齐序列的核的灵活性，而后者又取决于

7. “仅当”方向很直接：如果 k'_s 不是严格正定的，就会违背定理 13。

我们用于比较字母（包括实际字母，即核苷酸/氨基酸，以及间隙）的核的灵活性。替换参数与插入参数之间的相互作用，源于这样一个事实：我们实际上已将空位符号加入字母表，因此需要确保扩展后的字母表上的字母核仍然严格正定。

G. 比对核的特征

本节将推导比对核以特征表示的形式。这些特征是带间隙的 k-mer 计数。

定理 *40.*（定理 *26* 的证明）考虑一个比对核 k，其参数为 $k_s(b,b') = \sigma^{-1}|\mathcal{B}|\delta_{b'}(b)$，其中 $b, b' \in \mathcal{B}$。用 $A(X) = e^{\mu|X|}$ 对该核进行倾斜，得到 $k^A(X,Y) = A(X)A(Y)k(X,Y)$。对于所有序列 $X \in S$，定义特征 $\{u_V(X)\}_{V\in S}$，如下所示：

$$u_V(X) = e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g} \sum_{J\in G(|X|,g)} \mathbb{1}(X_{(J)} = V), \tag{13}$$

其中 $\zeta = 2\mu - \log\sigma + \log|\mathcal{B}|$。现在我们证明，$\{u_V(X)\}_{V\in S}$ 是 $\mathcal{H}_{k^A}$ 的标准正交基。

证明 我们可以用索引对 $J \in G(|X|, M)$ $J' \in G(|Y|, M)$ 表示两个序列 $X \in S$ 和 $Y \in S$ 之间的比对，这些索引对描述了 X 和 Y 中哪些字母相互比对。特别地，$k(X,Y)$ 等于

$$\sum_{M=0}^{\infty} \sum_{(j_l)_{l=-1}^{M}\in G(|X|,M)} \sum_{(j'_l)_{l=-1}^{M}\in G(|Y|,M)} k_i(X_{(:j_0)}, Y_{(:j'_0)}) \times \prod_{m=0}^{M-1} k_s(X_{(j_m)}, Y_{(j'_m)}) k_i(X_{(j_m+1:j_{m+1})}, Y_{(j'_m+1:j'_{m+1})})$$

根据我们对 k_s 的假设，对于每个比对 $J \in G(|X|, M, g)$ $J' \in G(|Y|, M, g')$，比对字母的贡献 $\prod_{m=0}^{M-1} k_s(X_{(j_m)}\ Y_{(j'_m)})$，如果 $X_{(J)} \neq Y_{(J')}$ 则为 0，否则为 $\sigma^{-M}|\mathcal{B}|^M$。另一方面，插入和缺失的贡献为 $\prod_{m=0}^{M-1} k_I(X_{(j_m:j_{m+1})}\ Y_{(j'_m+1:j'_{m+1})}) = e^{-\mu(|X|-M+|Y|-M)}e^{-\Delta\mu(g+g')}$。因此，$k^A(X,Y)$ 等于

$$\begin{aligned}
&\sum_{M=0}^{\infty} \left(e^{2\mu}\sigma^{-1}|\mathcal{B}|\right)^M \sum_{g,g'} e^{-\Delta\mu(g+g')} \sum_{J\in G(|X|,M,g)} \sum_{J'\in G(|Y|,M,g')} \mathbb{1}(X_{(J)} = Y_{(J')}) \\
&= \sum_{M=0}^{\infty} e^{\zeta M} \sum_{V\in\mathcal{B}^M} \sum_{g,g'} e^{-\Delta\mu(g+g')} \sum_{J\in G(|X|,M,g)} \sum_{J'\in G(|Y|,M,g')} \mathbb{1}(X_{(J)} = V)\mathbb{1}(Y_{(J')} = V) \\
&= \sum_{M=0}^{\infty} \sum_{V\in\mathcal{B}^M} u_V(X)u_V(Y)
\end{aligned} \tag{14}$$

因此，k^A 的特征为 $X \mapsto (u_V(X))_{V\in S}$。

现在我们将证明，对所有 $V \in S$， $u_V \in \mathcal{H}_k$ 都成立，并且 $(u_V)_V$ 构成一个标准正交基。将无限矩阵 Q 定义为 $Q_{V,\,Z} = e^{-\frac{1}{2}\zeta|V|}u_V(Z)$，对所有 V， $Z \in S$均成立。对 S 选取某个全序 $\leq$，使得若 $|V| \leq |Z|$，则 $V \leq Z$。在这种情况下， Q 是一个对角线上全为 1 的无限上三角矩阵。因此， Q 的逆矩阵是一个对角线上全为 1 的无限上三角矩阵 Q^{-1}。在证明的其余部分，我们将任意 $\sum_Z \alpha_Z k_Z^A \in$张成空间$\{k_Z^A\}_{Z\in S}$表示为一个无限向量，其索引为$S$的元素，且 α_Z 位于位置 Z。因此，对于任意函数 $f \in$ 张成空间$\{k_Z^A\}_{Z\in S}$，我们可以将矩阵向量乘积 Qf 视为 f的特征化。现在，对于任意 f， $g \in$张成空间$\{k_Z^A\}_{Z\in S}$，由公式 14 可得，

$$(f|g)_{k^A} = f^T Q^T \operatorname{diag}(e^{\zeta|V|})_{V\in S} Q g,$$

其中，$\operatorname{diag}(e^{\zeta|V|})_{V\in S}$表示一个对角线元素为 $e^{\zeta|V|}$的无限对角矩阵。

接下来我们将证明 $u_X \in \mathcal{H}_{k^A}$，其表征为 $f_X = e^{-\frac{1}{2}\zeta|X|}\sum_Z Q_{Z,\,X}^{-1} k_Z \in$张成空间 $\{k_Z^A\}_{Z\in S}$（注意，由于只有有限个 Z 满足 $Z \leq X$，因此该和只有有限个非零项）$Q_{Z,\,X}^{-1} \neq 0$。我们有，

$$\operatorname{diag}(e^{\zeta|V|})_{V\in S} Q f_X = e^{\frac{1}{2}\zeta|X|} d_X,$$

其中， d_X是一个以 S 为索引的无限维向量，除位置 X 上的值为 1 外，其余位置的值均为 0。对于任意 $Y \in S$，$(f_X|k_Y^A)_{k^A} = e^{\frac{1}{2}\zeta|X|}Q_{X,\,Y} = u_X(Y)$。因此， $u_X \in \mathcal{H}_{k^A}$。我们还有，

$$(u_X|u_Y)_{k^A} = f_X^T Q^T \operatorname{diag}(e^{\zeta|V|})_{V\in S} Q f_Y = d_X^T d_Y = d_X(Y).$$

因此， $\{u_X\}_{X\in S}$是由正交非零向量构成的集合，并经过归一化，使得对于所有 $X \in S$，$\|u_X\|_{k^A} = 1$。最后，根据公式 14，

$$k^A(X,Y) = \sum_{V\in S}^{\infty} (u_V|k_X^A)_{k^A}(u_V|k_Y^A)_{k^A}$$

对所有 X， $Y \in S$使得 $\{u_X\}_{X\in S}$也是一个基。 ■

现在我们考虑一般情形，允许 k_s 是字母上的任意严格正定核，而不要求它是对角的。第一步是证明，任意比对核都可以表示为一个具有对角 k_s 的比对核的重新参数化。

命题 41。考虑一个比对核 k，其参数为 μ、 $\Delta\mu$ 和 k_s。令$K_{b,\,b'} = k_s(b,b')$，其中 b、$b' \in \mathcal{B}$和 $\sigma = \mathbf{1}_{\mathcal{B}}^T K^{-1}\mathbf{1}_{\mathcal{B}}$。存在一个字母表 $\tilde{\mathcal{B}}$，它导出由序列 S构成的集合 $\tilde{\ }$ $\tilde{\ }$，使得将 k 应用于 S 后，得到一个参数为 μ、$\Delta\mu$ 和$k_s(\tilde{b},\tilde{b}') = \sigma^{-1}|\mathcal{B}|\delta_{\tilde{b}}(\tilde{b}')$ 的比对核，其中对所有 $\tilde{b}$、$\tilde{b}' \in \tilde{\mathcal{B}}$均成立。此外，对于任意 $b \in \mathcal{B}$、 $\sum_{\tilde{b}\in\tilde{\mathcal{B}}} k_s(\tilde{b},b) = \sigma^{-1}|\mathcal{B}|$。我们 $\tilde{\ }$ $\tilde{\ }$将 $\mathcal{B}$ 称为 k 的校正字母表，并将 S 称为 k的校正序列集。

证明 令 $B_1 = K^{-1}\mathbf{1}_{\mathcal{B}}/\sqrt{\sigma}$因此 $k_s(B_1, B_1) = 1$并令 $\{B_2,\ldots,B_{|\mathcal{B}|}\}$为与 B_1 正交的正交归一集合，位于 $\mathbb{R}^{\mathcal{B}}$中，内积为 k_s。令 $\Phi \in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{B}|}$ 为一个酉矩阵，其第一列的每个元素均为 $|\mathcal{B}|^{-1/2}$，并定义字母 $C_i = \sqrt{\sigma^{-1}|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\Phi_{i,\,j}B_j$。因此我们有 $k_s(C_i, C_j) = \sigma^{-1}|\mathcal{B}|(\Phi_{i,\,\cdot}|\Phi_{j,\,\cdot}) = \sigma^{-1}|\mathcal{B}|\delta_i(j)$ 且 $\sum_{b\in\mathcal{B}} C_{i,\,b} = C_i^T\mathbf{1}_{\mathcal{B}} = \sqrt{\sigma}C_i^T K B_1 = \sqrt{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\Phi_{i,\,j}k_s(B_i, B_1) = \sqrt{|\mathcal{B}|}\Phi_{i,\,1} = 1$。

称 $\mathcal{B} = \{C_i\}_i$ 和 S 为字母取自 $\mathcal{B}$的序列集合。根据命题 38，对于$V, V' \in S$, $k(V, V')$ 由给出

$$\sum_{l\in\mathbb{N}, V^{(1)}+\cdots+V^{(2l+1)}=V, V'^{(1)}+\cdots+V'^{(2l+1)}=V'} k_I(V^{(1)}, V'^{(1)}) \prod_{i=1}^{l} k_s(V^{(2i)}, V'^{(2i)}) k_I(V^{(2i+1)}, V'^{(2i+1)}).$$

对于 $\tilde{b}$, $\tilde{b}' \in \tilde{\mathcal{B}}$, $k_s(\tilde{b}, \tilde{b}') = \sigma^{-1}|\mathcal{B}|\delta_{\tilde{b}}(\tilde{b}')$。此外，对于任意数值 L和 $V \in \tilde{\mathcal{B}}^L$，作用于 V 的插入核 k_I 为，

$$k_{I,V} = \sum_{|X|=L} \left(\prod_{l=0}^{L-1} V_{l,X_{(l)}} \right) e^{-\Delta\mu - L\mu} g = e^{-\Delta\mu - L\mu} \prod_{l=0}^{L-1} \left(\sum_{b\in\mathcal{B}} V_{l,b} \right) g = e^{-\Delta\mu - L\mu} g.$$

因此，对于 V, $V' \in S$以及 $|V|$, $|V'| > 0$,

$$k_I(V, V') = e^{-\Delta\mu - |V|\mu} e^{-\Delta\mu - |V'|\mu}.$$

因此， k 在 S上的比对核具有相同的参数 μ, $\Delta\mu$，并且对于 $\tilde{b}$, $\tilde{b}' \in \tilde{\mathcal{B}}$, $k_s(b, b') = \sigma^{-1}|\mathcal{B}|\delta_{\tilde{b}}(\tilde{b}')$。

最后请注意，对于任意 $b \in \mathcal{B}$,

$$\begin{aligned}
\sum_i k_s(C_i, b) &= \sqrt{\sigma^{-1}|\mathcal{B}|} \sum_{i,j} \Phi_{i,j} k_s(B_j, b) \\
&= \sqrt{\sigma^{-1}|\mathcal{B}|} \sum_j \left(\mathbf{1}_{|\mathcal{B}|}^T \Phi_{\cdot,j} \right) k_s(B_j, b) \\
&= \sqrt{\sigma^{-1}|\mathcal{B}|} \sum_j \sqrt{|\mathcal{B}|}\delta_1(j) k_s(B_j, b) \\
&= \sqrt{\sigma^{-1}}|\mathcal{B}| k_s(B_1, b) \\
&= \sigma^{-1}|\mathcal{B}|.
\end{aligned}$$

■

现在我们可以描述具有任意 k_s 的比对核的特征。当 k_s 为对角矩阵时，序列 X 的每个特征都取决于间隔 k-mer 精确匹配的位置数量 X 。现在，我们不再要求精确匹配，而是根据 k_s为每次匹配评分。

推论 42。考虑一个比对核 k ，并令 S 为一个校正序列集。对于 $Y \in S$、 $V \in S$ 且满足 $L = |Y| = |V|$的情况，使用 k_s 定义成对比较如下，

$$c_{pw}(Y, V) = \prod_{l=0}^{L-1} \sigma|\mathcal{B}|^{-1} k_s(Y_{(l)}, V_{(l)}).$$

将核倾斜 $A(X) = e^{\mu|X|}$ 可得 $k^A(X, Y) = A(X)A(Y)k(X, Y)$。对于所有序列$X \in S$, 定义特征 $\{u_V(X)\}_{V\in\tilde{S}}$，其定义为

$$u_V(X) = e^{\frac{1}{2}\zeta|V|} \sum_{g=0}^{\infty} e^{-\Delta\mu g} \sum_{J\in G(|X|,g)} c_{pw}(X_{(J)}, V), \tag{15}$$

其中 $\zeta = 2\mu -$ 对数 $\sigma +$ 对数 $|\mathcal{B}|$。现在， $\{u_V(X)\}_{V\in\tilde{S}}$是 $\mathcal{H}_{k^A}$ 的标准正交基。

˜ ˜**证明**令 $\mathcal{B}$, S 一个校正字母表和一个校正序列集，对于 k。根据定理26，˜存在一个 $\mathcal{H}_{k^A}$的标准正交基， $\{u_V\}_{V\in\tilde{S}}$使得对于任意 $V, X \in S$

$$(u_V|k_X^A)_{k^A} = e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)} \mathbb{1}(X_{(J)} = V).$$

注意，重新参数化不会影响再生核希尔伯特空间 $\mathcal{H}_{k^A}$，因此 $\{u_V\}_{V\in\tilde{S}}$ 是应用于空间 S 的核 k^A 的标准正交基。

现在，考虑一个 $Y \in S$，其中 $L = |Y|$。由于 $\tilde{\mathcal{B}}$ 是 $\mathbb{R}^{\mathcal{B}}$ 的一组基，我们可以将每个字母˜$Y_{(l)}$写作 $Y_{(l)} = \sum_{\tilde{b}\in\tilde{\mathcal{B}}} Y_{(l),\ \tilde{b}}\, b$，其中系数 $\{Y_{(l),\tilde{b}}\}_{\tilde{b}\in\tilde{\mathcal{B}}}$ 为某些标量（说明一下，这里我们将$Y_{(l)}$ 视为 Y 中第 l 个字母的独热编码）。为推导系数 $Y_{(l),\ \tilde{b}}$，回想一下，$\{\sqrt{\sigma|\mathcal{B}|^{-1}}\,\tilde{b}\}_{\tilde{b}\in\tilde{\mathcal{B}}}$是 $\mathbb{R}^{\mathcal{B}}$ 的一组标准正交基，其内积定义为 k_s。因此，对于所有 $\tilde{b} \in \tilde{\mathcal{B}}$，有 $Y_{(l),\ \tilde{b}} = k_s(Y_{(l)},\ \sqrt{\sigma|\mathcal{B}|^{-1}}\,\tilde{b})\sqrt{\sigma|\mathcal{B}|^{-1}} = \sigma|\mathcal{B}|^{-1}k_s(Y_{(l)},\ \tilde{b})$。此外，根据命题 41 中校正字母表的性质，我们有 $\sum_{\tilde{b}\in\tilde{\mathcal{B}}} Y_{(l),\ \tilde{b}} = \sigma|\mathcal{B}|^{-1}\sum_{\tilde{b}\in\tilde{\mathcal{B}}} k_s(Y_{(l)},\ \tilde{b}) = 1$。因此，

$$\begin{aligned}
(u_V|k_Y^A)_k^A &= \sum_{X\in\tilde{S}\,:\,|X|=L}\left(\prod_{l=0}^{L-1} Y_{(l),X_{(l)}}\right) e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)} \mathbb{1}(X_{(J)} = V)\\
&= e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)}\ \sum_{X\in\tilde{S}\,:\,|X|=L}\left(\prod_{l=0}^{L-1} Y_{(l),X_{(l)}}\right)\prod_{m=0}^{M-1}\mathbb{1}(X_{(j_m)} = V_{(m)})\\
&= e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)}\left(\prod_{l\notin J}\sum_{b\in\tilde{\mathcal{B}}} Y_{(l),b}\right)\prod_{m=0}^{M-1} Y_{(j_m),V_{(m)}}\\
&= e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)} 1\times\prod_{m=0}^{M-1}\sigma|\mathcal{B}|^{-1}k_s(Y_{(j_m)},V_{(m)})\\
&= e^{\frac{1}{2}\zeta|V|}\sum_{g=0}^{\infty} e^{-\Delta\mu g}\sum_{J\in G(|X|,g)} c_{pw}(Y_{(J)},V)
\end{aligned}$$

■

H. 比对 核在某些倾斜下是通用且具有特征性的 s

在本节中，我们将证明，在适当的倾斜下，所有比对核都是通用且具有特征性的——即使它们没有离散质量。请注意，具有离散质量的核在任何倾斜后仍保持离散质量，但通用核或特征核并非如此；在这里，倾斜方式至关重要。

函数在 C_0 在无穷远处消失，因此，为确保比对核能够逼近任意 C_0 函数，即使该核没有离散质量，倾斜也必须足够小。归一化倾斜 $A(X) = \sqrt{k(X,X)}^{-1}$ 未必

足够小。相反，通常需要进行如此大的倾斜，使得 $k^A(X,X) \to 0$ 当$|X| \to \infty$。从建模角度来看，这会带来问题：如果 f 是从核为 k^A的高斯过程中抽取的，那么随着序列长度增加，其方差将趋于零：$\mathrm{Var}(f(X)) = k^A(X,X) \to 0$ 当 $|X| \to \infty$。因此，对于较长的序列， f 的值基本上只由先验决定。所以，尽管我们可以通过足够大的倾斜，使缺少离散质量的比对核成为通用核，但倾斜后的核泛化能力会很差。

命题 *43.* 设 k 是一个比对核，并令 $\tilde{k}$ 是一个比对核，它与 k 在 μ、$\Delta\mu$ 处的取值相同，但对于某个 $\epsilon > 0$，有 $\tilde{k}_s(b,b') = \sigma e^{-2\mu+\epsilon} k_s(b,b')$。取一个 $A: S \to (0,\infty)$，使得 $\tilde{k}^A$ 是一个 C_0-核，即对于每个 X， $Y \mapsto A(Y)A(X)\tilde{k}(X,Y)$ 属于 $C_0(S)$，且 $\sup_{X\in S} A(X)A(X)\tilde{k}(X,X) < \infty$。那么，倾斜比对核 k^A 是 C_0-通用且具有特征性的。

证明 以下始终假设 $2\mu \le \log\sigma$，否则结论可由定理 23 和推论 18 得出。注意，对所有 $X, Y \in S$， $\tilde{k}(X,Y) \ge k(X,Y)$。因此， k^A 是一个 C_0-核。另请注意，定义 $\tilde{K}_{b,b'} = \tilde{k}_s(b,b')$ 后，我们有 $\log(\mathbf{1}_{\mathcal{B}}\tilde{K}^{-1}\mathbf{1}_{\mathcal{B}}) = 2\mu - \epsilon < 2\mu$，因此根据定理 23， $\tilde{k}$ 具有离散质量。

我们现在推导出 $\mathcal{H}_{\tilde{k}^A}$ 的一个正交基。令 $A(X) = \exp(-\mu|X|)$，并令 S 为 k^A 的一个校正序列集。根据推论 42 和命题 35， $\{A\bar{A}u_V\}_{V\in\tilde{S}}$是 $\mathcal{H}_{k^A}$ 的一个标准正交基。注意， $\tilde{S}$ 也是 $\tilde{k}^A$ 的一个校正序列集，因此令$\{AA\tilde{u}_V\}_{V\in\tilde{S}}$为推论 42 给出的 $\mathcal{H}_{\tilde{k}^A}$ 的标准正交基。注意，推论 42 中的 c_{pw} 对于 k^A 和 $\tilde{k}^A$ 是相同的，因此每个 $\tilde{u}_V$ 都是 u_V 的一个缩放版本；特别地，$\{AAu_V\}_{V\in\tilde{S}}$是 $\mathcal{H}_{\tilde{k}^A}$ 的一个正交基。

由于 k 具有离散质量，因此它是通用的，所以对于任意 $f \in C_0(S)$ 和 $\epsilon > 0$，我们都可以选取一个 $g \in \mathcal{H}_{\tilde{k}}$，使得 $\|f-g\|_\infty < \epsilon$。定义 $M = \sup_{X\in S} \sqrt{A(X)A(X)\tilde{k}(X,X)} < \infty$，并选取一个 $g' \in$张成空间$\{\tilde{k}^A_X\}_X$，使得 $\|g'-g\|_{\tilde{k}^A} \le \epsilon/M$。现在， $g' \in C_0(S)$，且

$$|g'(X) - g(X)| = |(\tilde{k}^A_X | g'-g)_{\tilde{k}^A}| \le \|\tilde{k}^A_X\|_{\tilde{k}^A}\|g'-g\|_{\tilde{k}^A} \le M\epsilon/M = \epsilon$$

对于所有 $X \in S$。因此， $\|g'-f\|_\infty \le 2\epsilon$。因此， $\mathcal{H}_{\tilde{k}^A}$ 中的每个函数都可以用 $\{\tilde{k}^A_X\}_X$ 的张成空间中的函数在C_0中逼近。因此，$\{\tilde{k}^A_X\}_X$的张成空间在 $C_0(S)$ 中稠密。

注意有限维向量空间 $\mathrm{span}\{AAu_X\}_{X\in\tilde{S}:\ |X|\le L}$对所有 L都相同（回想命题 19 的证明，重新参数化字母表会保留核嵌入向量的张成空间）。因此，$\mathrm{span}\{\tilde{k}^A_X\}_{X\in S} = \mathrm{span}\{A\bar{A}u_X\}_{X\in\tilde{S}}$。但后一个向量空间同样等于 $\mathrm{span}\{k^A_X\}_X$。由此，$\mathrm{span}\{k^A_X\}_X = \mathrm{span}\{\tilde{k}^A_X\}_X$，所以 $\mathrm{span}\{k^A_X\}_X$在$C_0(S)$ 中稠密。 k^A 因此是通用的。此外，根据 Sriperumbudur 等人（2011）的命题 2， k^A也具有特征性。

■

为了解读这一结果及其所需的倾斜，请注意，我们必须选择一种倾斜 A，使得 $\tilde{k}^A(X,X)$ 有界。粗略地说，我们预计 $k(X,X)$ 会随着长度 $|X|$呈指数增长，因为每次成对比对中的每个匹配都被赋予了 $\sigma e^{-2\mu+\epsilon} > 1$倍的权重。归一化倾斜 $A(X) = \sqrt{k(X,X)}^{-1}$ 将有助于我们界定

$k^A(X, X)$，但要界定 $\tilde{k}^A(X, X)$，我们还需要进一步降低较长序列的权重，且降低幅度呈指数级。这意味着要选择一种倾斜 A，使得 $k^A(X, X) \to 0$ 随着 $|X| \to 0$。

I. 单字母比对核中的离散质量

本节将证明，在某些参数设置下，当字母表大小为一时，比对核具有离散质量。这一结果对于证明适用于任意字母表大小的更一般结果至关重要，相关内容见定理 23 和附录 E。该结果建立在定理 26 和附录 G 中所述的特征表示之上。

该结果将依赖 Riordan 群理论，因此我们需要先引入一些记号。若 $f(x) = \sum_{i=0}^{\infty} C_i x^i$ 为形式幂级数，则定义 $[x^n]f = C_n$。对于两个形式幂级数 $f(y)$ $g(y)$，我们将 f 和 g 的 Riordan 阵列定义为一个无限上三角矩阵，其元素为

$$Q_{L',L} = [x^{L'} y^L] \frac{g(y)}{1 - xyf(y)} = [x^{L'} y^L] \sum_{l=0}^{\infty} g(y) x^l y^l f(y)^l = [y^L] y^{L'} g(y) f(y)^{L'}$$

对于数 L，$L' \geq 0$。若 $f = g$，则称 Q 属于 Bell 子群。我们需要的结果是，可以通过对幂级数 f 和 g 求逆来求 Riordan 阵列的逆（Shapiro 等，1991）。

命题 44。 （*Shapiro* 等，*1991*）设

$$Q_{L',L} = [x^{L'} y^L] \frac{g(y)}{1 - xyf(y)}$$

并令 $\bar{f}$ 为满足方程 $y\bar{f}(y)f(y\bar{f}(y)) = y$ 的唯一幂级数。[8] 然后，令 $\bar{g}(y) = \left(g(y\bar{f}(y))\right)^{-1}$，

$$Q^{-1}_{L',L} = [x^{L'} y^L] \frac{\bar{g}(y)}{1 - xy\bar{f}(y)}.$$

注意，如果 Q 属于 *Bell* 子群，那么 $g(y) = f(y)$ 和 $f(y)f(yf(y)) = 1$，因此 $\bar{g}(y) = f(y)$。由此可知，Q^{-1} 也属于 *Bell* 子群。

证明

$$\begin{aligned}
\sum_{l=0}^{\infty} Q_{L,l} Q^{-1}_{l,L'} &= \sum_{l=0}^{\infty} [y^l] g(y) \left(yf(y)\right)^L [z^{L'}] \bar{g}(z) \left(z\bar{f}(z)\right)^l \\
&= [z^{L'}] \bar{g}(z) \sum_{l=0}^{\infty} \left(z\bar{f}(z)\right)^l [y^l] g(y) \left(yf(y)\right)^L \\
&= [z^{L'}] \bar{g}(z) g(z\bar{f}(z)) \left(z\bar{f}(z) f(z\bar{f}(z))\right)^L \\
&= [z^{L'}] 1 \times z^L = \delta_{L'}(L).
\end{aligned}$$

■

8. 该条件的另一种写法是 $tf(t) = y$ 当 $t = y\bar{f}(y)$，因此 yf 是 $y\bar{f}$ 的复合逆矩阵。

定理 45. 设 $|\mathcal{B}| = 1$，并称 A 为 $\mathcal{B}$ 中唯一的字母。若 $\Delta\mu > 0$，则 k 具有离散质量，当且仅当 $2\mu +$ 对数 $k_s(A, A) \geq 0$。若 $\Delta\mu = 0$，则 k 具有离散质量，当且仅当$2\mu +$ 对数 $k_s(A, A) > 0$。

证明 注意，在这种情况下 $\zeta = 2\mu + \log k_s(A, A)$。我们将在全文中用 $k^A(X, Y) = e^{\mu|X|}e^{\mu|Y|}k(X, Y)$ 替换 k。根据命题 35，这不会影响该核是否具有离散质量。索引 $S = \{0, 1, \dots\}$。将 Q 称为无限上三角矩阵，其中 $e^{-\frac{1}{2}\zeta L'}u_{L'}(L) = \sum_{g=0}^{\infty} e^{-g\Delta\mu}|G(L, L', g)|$ 位于位置 (L', L)。注意，k 在 $\{0, 1, \dots L\}$ 上的格拉姆矩阵为 $Q^T_{:L+1,:L+1}\mathrm{diag}(e^{\zeta L'})_{L'=0}^{L}Q_{:L+1,:L+1}$。由于 Q 是对角线上没有零元素的无限上三角矩阵，因此它具有无限上三角逆矩阵Q^{-1}，且满足 $(Q_{:L,:L})^{-1} = (Q^{-1})_{:L,:L}$。因此，$k$在 $\{0, 1, \dots L\}$ 上的格拉姆矩阵的逆矩阵为 $Q^{-1T}_{:L+1,:L+1}\mathrm{diag}(e^{-\zeta L'})_{L'=0}^{L}Q^{-1}_{:L+1,:L+1}$，其在位置 (L', L') 上的元素为 $\sum_{l=L'}^{L} e^{-\zeta l}(Q^{-1}_{L',l})^2$。因此，根据命题 13，$\delta_{L'} \in \mathcal{H}_k$ 当且仅当

$$\sum_{l=L'}^{\infty} e^{-\zeta l}(Q^{-1}_{L',l})^2 < \infty. \tag{16}$$

接下来将分两部分运用生成函数技巧和 Riordan 群理论来进行证明。第一部分，我们将 $u_L(L')$与某个形式幂级数的系数对应起来，然后通过求该幂级数的逆，显式求出矩阵 Q 的逆。第二部分，我们将方程 16 与幂级数的收敛性对应起来，并利用这一点证明结论。

第一部分：设 g 是大于或等于 1 的整数，且 $L \geq L'$。我们要找出 $G(L, L', g)$ 的大小。每个 $J = (j_{-1}, \dots, j_{L'}) \in G(L, L', g)$ 都由数值 $j_{m_1}, \dots, j_{m_g}$唯一确定，这些数值表示每个 g 间隙开始的位置；而$j_{m_1+1}, \dots, j_{m_g+1}$表示每个 g 间隙结束的位置。特别地，

$$J = (-1, 0, 1, \dots, j_{m_1} - 1, j_{m_1}, j_{m_1+1}, j_{m_1+1} + 1, \dots, j_{m_2} - 1, j_{m_2}, j_{m_2+1}, \dots, L).$$

因此，每个 J 都可以由其 g 间隙 $l_1 = j_{m_1+1} - j_{m_1} - 1, l_2 = j_{m_2+1} - j_{m_2} - 1, \dots$ 的长度以及其 $g + 1$ 连续区域 $c_0 = j_{m_1} + 1, c_1 = j_{m_2} - j_{m_1+1} + 1, c_2 = j_{m_3} - j_{m_2+1} + 1, \dots,$ $c_g = L - j_{m_g+1}$ 的长度唯一确定。另一方面，给定数值 $l_1, \dots, l_g, c_0, \dots, c_g$，当且仅当对于所有 i 都有 $\sum_{i=0}^{g} c_i = L', \sum_{i=1}^{g} l_i = L - L', l_i \geq 1$，且对于所有 $i = 1, 2, \dots, g - 1$ 都有 $c_i \geq 1$ 时，便可得到唯一有效的 $J \in G(L, L', g)$（$c_0 = 0$ 表示序列开头有一个间隙，而 $c_g = 0$ 表示序列末尾有一个间隙）。满足 $\sum_{i=0}^{g} c_i = L'$ 且对于所有 $i = 1, 2, \dots, g - 1$ 都有 $c_i \geq 1$ 的 $g + 1$ 个有序数值 $(c_0, \dots, c_g)$ 的数量，是该幂级数的第 L' 个系数

$$(1 + x + x^2 + \dots)\left(x + x^2 + \dots\right)^{g-1}(1 + x + x^2 + \dots) = \frac{x^{g-1}}{(1 - x)^{g+1}}.$$

类似地，对于所有 i，满足 $\sum_{i=1}^{g} l_i = L - L'$ 且 $l_i \geq 1$ 的 g 有序数列 $(l_1, \dots, l_g)$ 的集合数量，是幂级数 $\frac{y^g}{(1-y)^g}$ 的 $L - L'$ 次系数。另一方面，如果 $g = 0$，那么 $G(L, L', g)$ 为空；否则它恰有一个元素。特别地，$G(L, L', 0)$ 是 $1/(1 - x)$ 中 $x^{L'}y^{L-L'}$ 前面的系数（显然如此）。因此，其大小为

$G(L,\ L',\ g)$ 是幂级数中 $x^{L'}y^{L-L'}z^g$ 前面的系数

$$\begin{aligned}
\sum_{g=1}^{\infty} z^g \frac{x^{g-1}}{(1-x)^{g+1}} \frac{y^g}{(1-y)^g} + \frac{1}{1-x} &= \frac{zy}{(1-x)^2(1-y)} \sum_{g=0}^{\infty} \left(z \frac{x}{1-x} \frac{y}{1-y} \right)^g + \frac{1}{1-x} \\
&= \frac{zy}{(1-x)^2(1-y)} \frac{(1-x)(1-y)}{(1-x)(1-y)-zxy} + \frac{1}{1-x} \\
&= \frac{1}{1-x} \left(\frac{zy}{(1-x)(1-y)-zxy} + 1 \right) \\
&= \frac{1}{1-x} \left(\frac{zy+1-x-y+xy-zxy}{(1-x)(1-y)-zxy} \right) \\
&= \frac{1}{1-x} \left(\frac{(zy+1-y)(1-x)}{(1-x)(1-y)-zxy} \right) \\
&= \frac{1-(1-z)y}{(1-x)(1-y)-zxy} \\
&= \frac{\frac{1-(1-z)y}{1-y}}{1-x\frac{1-(1-z)y}{1-y}}.
\end{aligned}$$

称 $f_z(y) = \frac{1-\frac{1-(1-z)y}{1-y}}{y}$。接下来我们用 xy 替换 x，从而得到 $|G(L,L',g)|$ 是 $[(xy)^{L'}y^{L-L'}z^g] f_{z(y)\,1-(xy)f_z(y)} = [x^{L'}y^L z^g] \frac{f_z}{(y)\,1-xyf_z(y)}$。因此，代入 $z = e^{-\Delta\mu}$ 并记 $f = f_{e^{-\Delta\mu}}$,

$$Q_{L',L} = \sum_{g=0}^{\infty} e^{-g\Delta\mu} |G(L,L',g)| = [x^{L'}y^L] \sum_{g=0}^{\infty} e^{-g\Delta\mu} [z^g] \frac{f_z(y)}{1-xyf_z(y)} = [x^{L'}y^L] \frac{f(y)}{1-xyf(y)}.$$

Q 属于 Riordan 数组的Bell 子群（Shapiro等人（1991 年）以及Shapiro等人（2022年）的第3 章）。这意味着，根据命题44，将 f 定义为具有 $yf(y)f(yf(y)) = y$的唯一幂级数，

$$Q^{-1}_{L',L} = [x^{L'}y^L] \frac{\bar{f}(y)}{1-xy\bar{f}(y)}.$$

第 2 部分：由于

$$Q^{-1}_{L',L} = [x^{L'}y^L] \frac{\bar{f}(y)}{1-xy\bar{f}(y)} = [y^L]\bar{f}(y)(y\bar{f}(y))^{L'} = [y^{L-L'}]\bar{f}(y)^{L'+1},$$

我们有,

$$\sum_{l=L}^{\infty} e^{-\zeta l}(Q^{-1}_{L,l})^2 = e^{-\zeta L} \sum_{l=0}^{\infty} (e^{-\frac{1}{2}\zeta l}[y^l]\bar{f}(y)^{L+1})^2.$$

当 $e^{-\frac{1}{2}\zeta}$ 小于 $\bar{f}(y)^{L+1}$的收敛半径时，这个和是有限的（因为级数中的项呈指数衰减）；当它大于收敛半径时，这个和是无限的（因为各项不会趋近于 0）。现在我们将证明，f 的收敛半径恰好为 1，因此，若 $\zeta < 0$，则 $\delta_0 \notin \mathcal{H}_k$；若 $\zeta > 0$，则由于对于所有 L，$\bar{f}(y)^L$ 的收敛半径也至少为 1，k 具有离散质量。之后我们将仔细考察 $\zeta = 0$ 的情形。

我们先考虑 $\Delta\mu = 0$ 的情形。在这种情况下，根据 f 的定义，我们有 $yf(y) = \frac{y}{1-y}$，因此 $yf(y) = \frac{y}{1+y}$。显然，此时 f 的收敛半径为 1。我们也可以明确看出，$[y^l]\bar{f}(y)^{L+1} = (-1)^l\binom{L+l}{l}$。因此，当 $\zeta = 0$ 时，对于任何 L，方程 16 也不成立。

现在假设 $\Delta\mu > 0$。我们将推导 $\bar{f}$ 的形式。令 $\xi = 1 - e^{-\Delta\mu} \in (0, 1)$，且 $t = yf(y)$。

$$tf(t) = y \iff t(1-\xi t) = (1-t)y \iff 0 = \xi t^2 - (y+1)t + y.$$

因此，取在 $y = 0$ 时为 0 的那个根，我们得到

$$y\bar{f}(y) = \frac{1}{2\xi}\left(1 + y - \sqrt{(1+y)^2 - 4\xi y^2}\right).$$

多项式 $(1+y)^2 - 4\xi y^2$ 的根为

$$(2\xi - 1) \pm \sqrt{(2\xi-1)^2 - 1} = (2\xi - 1) \pm i\sqrt{1 - (2\xi-1)^2}.$$

注意，这两个根的范数均为 1，因此可写为 $e^{i\theta}, e^{-i\theta}$，其中 $\pi > \theta > 0$。因此，我们可以写成

$$y\bar{f}(y) = \frac{1}{2\xi}\left(1 + y - \sqrt{(1 - e^{i\theta}y)(1 - e^{-i\theta}y)}\right).$$

$\sqrt{1 - e^{i\theta}y}, \sqrt{1 - e^{-i\theta}y}$ 的收敛半径均为 1，因此 $y\bar{f}$ 的收敛半径至少为 1。另一方面，当 y 趋近于 $e^{i\theta}$ 或 $e^{-i\theta}$ 时，$y\bar{f}$ 的导数发散，因此 yf，从而 f 的收敛半径恰为 1。

$\zeta = 0$ 时的情况更为微妙。首先记作 $g(y) = \sqrt{(1+y)^2 - 4\xi y^2} = \sqrt{(1 - e^{i\theta}y)(1 - e^{-i\theta}y)}$。对于实数 $0 < r < 1$，定义

$$\begin{aligned}
q_L(r) &= \int_0^1 dt g(\sqrt{r}e^{2\pi it})^L g(\sqrt{r}e^{-2\pi it})^L \\
&= \int_0^1 dt \sum_n r^{n/2} e^{2\pi nit}[y^n]g(y)^L \sum_m r^{m/2}e^{-2\pi mit}[z^m]g(z)^L \\
&= \sum_{n,m}\left([y^n]g(y)^L[z^m]g(z)^L\right) r^{(n+m)/2}\int_0^1 dt e^{2\pi it(n-m)} \\
&= \sum_n ([y^n]g(y)^L)^2 r^n.
\end{aligned}$$

g 在 0 处解析且具有实系数，因此对于所有 L, n，$[y^n]g(y)^L)^2 \geq 0$。因此，根据单调收敛定理，$\sum_n([y^n]g(y)^L)^2 = \lim_{r\to 1^-} q_L(r)$。然而，

$$g(\sqrt{r}e^{2\pi iz})^L g(\sqrt{r}e^{-2\pi iz})^L = \left|1 - e^{i(\theta+2\pi z)}\sqrt{r}\right|^L \left|1 - e^{i(\theta-2\pi z)}\sqrt{r}\right|^L \leq 4^L,$$

因此，对于所有 L，有 $\sum_n([y^n]g(y)^L)^2<\infty$。现在注意

$$\begin{aligned}\sum_n([y^n](y\bar{f}(y))^L)^2 &= \sum_n \frac{1}{(2\xi)^{2L}}\left(\sum_{l=0}^{L}[y^n](1+y)^l(-1)^{L-l}g(y)^{L-l}\right)^2\\ &= \sum_n \frac{1}{(2\xi)^{2L}}\left(\sum_{l=0}^{L}(-1)^{L-l}\sum_{l'=0}^{l}\binom{l}{l'}[y^{n-l'}]g(y)^{L-l}\right)^2\\ &\le \sum_n \frac{1}{(2\xi)^{2L}}\left(L^2\max_{l,l'}\left|\binom{l}{l'}[y^{n-l'}]g(y)^{L-l}\right|\right)^2\\ &\le \frac{1}{(2\xi)^{2L}}(L^2)^2\sum_n\max_{l,l'}\left|\binom{l}{l'}^2\left([y^{n-l'}]g(y)^{L-l}\right)^2\right|\\ &\le \frac{1}{(2\xi)^{2L}}(L^2)^2\sum_{l=0}^{L}\sum_{l'=0}^{l}\binom{l}{l'}^2\sum_n\left([y^{n-l'}]g(y)^{L-l}\right)^2<\infty.\end{aligned}$$

因此，当 $\zeta=0$ 时，k 也具有离散质量。

■

J. 局部比对核中的离散质量

在本节中，我们证明在某些参数设置下，局部比对核具有离散质量。

定理 46。（定理 *25* 的证明）如果 $\Delta\mu=0$，则 k_{la} 具有离散质量，当且仅当 $2\mu>\log\sigma$。如果 $\infty>\Delta\mu>0$，则 k_{la} 具有离散质量，当且仅当 $2\mu\ge\log\sigma$。如果 $\Delta\mu=\infty$，则 k_{la} 具有离散质量。

证明 我们将仅给出证明的概要，着重说明其与常规序列核情形的差异。具体细节与定理 23 和定理 45 的证明非常相似。

首先，我们重复定理 23 的推理，将问题化简为 $|\mathcal{B}|=1$ 的情形。令 $U,W_V,V[t],\tilde{B}$ 如定理 23 中所述。首先注意，如果 $V=\emptyset$, W_V 只是定义在具有 $|\mathcal{B}|=1$ 的空间上的 k_{la}，并带有 $k_s(U,U)=\sigma^{-1}$，那么.....因此，如果这个局部比对核没有离散质量，那么 k_{la} 在整个空间上也没有。接下来考虑 $|V|\ge 1$，并定义左局部比对核

$$k_{\mathrm{l}}(X,Y)=\sum_{l=0}^{\infty}\tilde{k}_I\star(k_s\star k_I)^l$$

and 类似地定义右局部对齐核 k_{r} 。现在我们重复公式 1 的推理 2:

$$
\begin{aligned}
k(V[t], V[t']) = & \sum_{\substack{l\in\mathbb{N} \\ Y^{(1)}+\cdots+Y^{(2l+1)}=t_0\times U \\ Z^{(1)}+\cdots+Z^{(2l+1)}=t'_0\times U}} \sum_{\substack{l'\in\mathbb{N} \\ \tilde{Y}^{(1)}+\cdots+\tilde{Y}^{(2l'+1)}=V[t_{1:}] \\ \tilde{Z}^{(1)}+\cdots+\tilde{Z}^{(2l'+1)}=V[t'_{1:}]}} \Bigg(\\
& \times \left(\tilde{k}_I(Y^{(1)}, Z^{(1)}) \prod_{i=1}^{l} k_s(Y^{(2i)}, Z^{(2i)}) k_I(Y(2i+1), Z^{(2i+1)}) \right) \\
& \times k_s(V_{(0)}, V_{(0)}) \\
& \times \left(\prod_{i=1}^{l'} k_I(\tilde{Y}^{(2i-1)}, \tilde{Z}^{(2i-1)} k_s(\tilde{Y}^{(2i)}, \tilde{Z}^{(2i)})) \right) \times \tilde{k}_I(\tilde{Y}^{(1)}, \tilde{Z}^{(1)}) \Bigg) \\
= & k_{\mathrm{l}}(t_0 \times U, t'_0 \times U) k_{\mathrm{r}}(V[t_{1:}], V[t'_{1:}]) \\
= & \dots \\
= & k_{\mathrm{l}}(t_0 \times U, t'_0 \times U) \left(\prod_{l=1}^{L-1} k_{\mathrm{ali}}(t_l \times U, t'_l \times U) \right) k_{\mathrm{r}}(t_L \times U, t'_L \times U).
\end{aligned}
\tag{17}
$$

因此，如果 k_{ali}，k_{l}，和 k_{r} 满足 $k_s(U, U)$，且 $=\sigma^{-1}$ 和 $|\mathcal{B}| = 1$ 具有离散质量，那么 k_{la} 在 W_V 上也具有离散质量。因此，只要证明 $|\mathcal{B}| = 1$ 且 k_{ali}，k_{la}，k_{l}，和 k_{r} 的情形，定理即成立。

因此，我们研究 $|\mathcal{B}| = 1$ 的情形。我们已经证明了常规对齐核的定理。现在考虑 k_{l}。设 $G(L, L', g)$ 表示长度为 L 、含有 g 个间隙的序列中长度为 L' 的子串数量（**不包括左端的间隙**）。在这种情况下，$G(L, L', g)$ 的每个元素由其 $g+1$ 个连续区域的长度确定；除最后一个间隙之后的区域外，每个区域的长度都必须至少为 1；此外还有 $g+1$ 个间隙区域，除第一个间隙外，每个间隙区域的长度都必须至少为 1。因此，按照定理 45 中相同的论证，我们得到 $|G(L, L', g)|$ 是幂级数中 $x^{L'} y^{L-L'} z^g$ 前面的系数

$$
\begin{aligned}
\sum_{g=0}^{\infty} z^g \frac{x^g}{(1-x)^{g+1}} \frac{y^g}{(1-y)^{g+1}} = & \frac{1}{(1-x)(1-y)} \frac{(1-x)(1-y)}{(1-x)(1-y) - xyz} \\
= & \frac{\frac{1}{1-y}}{1 - x f_z(y)}.
\end{aligned}
$$

因此，在这种情况下，将矩阵 Q 定义为定理 45 中所述，

$$
Q_{L',L} = [x^{L'} y^L] \frac{\frac{1}{1-y}}{1 - xy f(y)}.
$$

我们可以利用命题 44 对该矩阵求逆。具体而言，与矩阵逆对应的幂级数的分子 $\bar{g}(y)$ 为

$$
\left(\left(\frac{1}{1-y} \right) \circ (y \bar{f}(y)) \right)^{-1} = 1 - y \bar{f}(y).
$$

因此，

$$Q_{L',L}^{-1} = [x^{L'}y^{L}]\frac{1-y\bar{f}(y)}{1-xy\bar{f}(y)}.$$

在这种情况下， k_{l} 当且仅当对所有 L都具有离散质量，

$$\sum_{l=0}^{\infty} e^{-\zeta l}(Q_{L,l}^{-1})^2 = \sum_{l=0}^{\infty}\left(e^{-\frac{1}{2}\zeta l}[y^l](1-y\bar{f}(y))(y\bar{f}(y))^L\right)^2 < \infty.$$

在 $\Delta\mu < \infty$的情况下，该定理的证明几乎完全相同。如果 $\Delta\mu = \infty$ ，那么$f(y) = 1$，因此 $\bar{f}(y) = 1$，并且$Q_{L',\ L}^{-1} = \delta_{L'}(L) - \delta_{L'+1}(L)$。因此，只有有限多个求和项非零，方程 16 得以满足。对于 k_{r}，情况也完全相同。

最后，我们来看 k_{la}。设 $G(L, L', g)$表示一个长度为 L 、包含 g 个间隙（**不包括任一端的间隙**）的序列中长度为 L' 的子串数量。在这种情况下，当$g > 0$时， $G(L, L', g)$的每个元素都由其 $g+1$ 个连续区域的长度确定，每个区域的长度至少为 1；并且有 $g+2$ 个间隙区域，每个间隙区域的长度至少为 1，但第一个和最后一个间隙除外。当 $g = 0$ 时，我们必须确保包含 $L' = 0$的情形，尽管当 $g > 0$时 $L' = 0$ 不可能。因此， $|G(L, L', g)|$ 是幂级数中 $x^{L'}y^{L-L'}z^{g}$ 前面的系数

$$\begin{aligned}\sum_{g=0}^{\infty} z^g \frac{x^{g+1}}{(1-x)^{g+1}}\frac{y^g}{(1-y)^{g+2}} + \frac{1}{1-y} =& \frac{x}{(1-x)(1-y)^2}\frac{(1-x)(1-y)}{(1-x)(1-y)-xyz} + \frac{1}{1-y}\\ =& x\frac{\left(\frac{1}{1-y}\right)^2}{1-xf_z(y)} + \frac{1}{1-y}.\end{aligned}$$

因此在这种情况下，

$$Q_{L',L} = [x^{L'}y^{L}]xy\frac{\left(\frac{1}{1-y}\right)^2}{1-xyf(y)} + \frac{1}{1-y}.$$

Q 的第一行只是 1，并且

$$(Q_{1:,1:})_{L',L} = [x^{L'}y^{L}]\frac{\left(\frac{1}{1-y}\right)^2}{1-xyf(y)}.$$

$Q_{1:,\ 1:}$ 是一个里奥丹阵列，其逆矩阵

$$(Q_{1:,1:})_{L',L}^{-1} = [x^{L'}y^{L}]\frac{\left(1-y\bar{f}(y)\right)^2}{1-xy\bar{f}(y)}.$$

令 R 为上三角矩阵，其元素为 $R_{1:,1:} = (Q_{1:,1:})^{-1}$, $R_{0,0} = 1$ ，且 $R_{0,1:} = -\mathbf{1}^T(Q_{1:,1:})^{-1}$ ，其中 $\mathbf{1}$ 为全 1 无限向量。我们将看到 $R = Q^{-1}$。首先，对于 $L, L' \geq 1$,

$$Q_{L,\cdot}R_{\cdot,L'} = \delta_L(L').$$

接下来，对于 $L \geq 0$,

$$Q_{L,\cdot}R_{\cdot,0} = \delta_L(0).$$

最后，对于 $L' \geq 1$,

$$Q_{0,\cdot} R_{\cdot,L'} = -(\mathbf{1}^T (Q_{1:,1:})^{-1})_{L'-1} + \mathbf{1}^T (Q_{1:,1:})^{-1}_{\cdot,L'-1} = 0.$$

因此 $R = Q^{-1}$。

由于 $\mathbf{1}_L = 1 = [x^L](1-x)^{-1}$,

$$\begin{aligned}
\left(\mathbf{1}^T (Q_{1:,1:})^{-1}\right)_L &= \sum_{L'} (Q_{1:,1:})^{-1}_{L',L} [x^{L'}](1-x)^{-1} \\
&= \sum_{L'} [y^L] \left((1 - y\bar{f}(y))^2 \left(y\bar{f}(y)\right)^{L'} \right) [x^{L'}](1-x)^{-1} \\
&= [y^L](1 - y\bar{f}(y))^2 \sum_{L'} \left(y\bar{f}(y)\right)^{L'} [x^{L'}](1-x)^{-1} \\
&= [y^L](1 - y\bar{f}(y))^2 (1 - y\bar{f}(y))^{-1} \\
&= [y^L](1 - y\bar{f}(y))
\end{aligned}$$

因此,

$$Q^{-1}_{L',L} = [x^{L'} y^L] xy \frac{\left(1 - y\bar{f}(y)\right)^2}{1 - xy\bar{f}(y)} + 1 - y(1 - y\bar{f}(y)).$$

当 $\Delta\mu < \infty$ 时，证明的其余部分与前一种情况几乎完全相同。当$\Delta\mu = \infty$, $Q_0^{-1}{}_{,L} = \delta_0(L) - \delta_1(L) + \delta_2(L)$ 且

$$(Q_{1:,1:})^{-1}_{L',L} = [x^{L'} y^L] \frac{(1-y)^2}{1-xy} = \delta_{L'}(L) - 2\delta_{L'+1}(L) + \delta_{L'+2}(L).$$

因此 k_{la} 再次根据方程 16 具有离散质量。

■

K. 厚尾比对核的计算

现在描述可用于计算例 10 和例 11 中所述厚尾比对核的动态规划过程。我们将把这两个例子都作为同一通用动态规划算法的特例来处理。设 $\ell(b, b')$ 是一个取值范围为 1,0 的函数，用于统计某类字母对的出现次数。若 $\ell(b, b') = 1$，则称配对 (b, b') 为“ℓ类”。

定义

$$M(i,j,l) = \sum^{(l)} \prod_{i=0}^{c-1} k_I(X^{(2i)}_{(:i)}, Y^{(2i)}_{(:j)}) k_s(X^{(2i+1)}_{(:i)}, Y^{(2i+1)}_{(:j)})$$

其中，求和遍历 c 以及所有满足 $l = \sum_{i=1}^{c} \ell(X^{(2i+1)}(:i)\ Y(2i+1)\ (:J)\ )$ 的划分 $X^{(0)}(:i) + \cdots + X^{(2c+1)}(:i) = X_{(:i)}$, $Y(0)\ (:j) + \cdots + Y(2c+1)\ (:j) = Y_{(:j)}$。因此，$M(i,j,l)$ 对 $X_{(:i)}$ 与 $Y_{(:j)}$ 之间所有具有 l 个 ℓ型比较、且以 $X_{(i-1)}$ 与 $Y_{(j-1)}$之间的匹配结束的比对求和。类似地，定义 $I_X(i,j,l)$，它是 $X_{(:i)}$

与 $Y_{(:j)}$ 的比对得分之和，这些比对具有 l 个 ℓ型比较，并以针对 X 而非 Y的插入结束。最后，类似地定义 $I_Y(i,j,l)$，它是 $X_{(:i)}$ 与 $Y_{(:j)}$的比对得分之和，这些比对具有 l 个 ℓ型比较，并以针对 Y的插入结束。将 $X_{(:i)}$ 与 $Y_{(:j)}$ 之间所有具有 l 个 ℓ型比较的比对之和称为，

$$J(i,j,l) = M(i,j,l) + I_X(i,j,l) + I_Y(i,j,l).$$

现在我们可以使用动态规划计算 M, I_X 和 I_Y。我们从以下位置初始化，

$$\begin{aligned} M(i,0,l) &= M(0,j,l) = 0 \text{ for } l,i,j \\ M(0,0,0) &= 1 \\ I_X(0,j,l) &= I_Y(i,0,l) = 0 \text{ for } l,i,j. \end{aligned}$$

更新方程为，

$$\begin{aligned} M(i,j,l) &= \mathbb{1}(l>0)\ell(X_{(i-1)},Y_{(j-1)})k_s(X_{(i-1)},Y_{(j-1)})J(i-1,j-1,l-1) \\ &\quad + \big(1-\ell(X_{(i-1)},Y_{(j-1)})\big)k_s(X_{(i-1)},Y_{(j-1)})J(i-1,j-1,l) \\ I_X(i,j,l) &= e^{-\Delta\mu-\mu}M(i-1,j,l) + e^{-\mu}I_X(i-1,j,l) \\ I_Y(i,j,l) &= e^{-\Delta\mu-\mu}M(i,j-1,l) + e^{-\Delta\mu-\mu}I_X(i,j-1,l) + e^{-\mu}I_Y(i-1,j,l). \end{aligned}$$

（我们采用如下任意约定：在任意两个相邻的匹配位置之间，Y中的插入位于 X中的插入之前；因此，I_Y不会出现在 I_X的更新方程中。）接下来，我们可以将 $R(L) = J(|X|, |Y|, L)$ 定义为 X和 Y所有具有 ℓ类 L 比较的比对之和。动态规划算法通过 $O(|X||Y|(|X| \wedge |Y|))$ 次运算计算 $R(L)$。

现在考虑选择 $k_s(b, b') = 1$ 和 $\ell(b, b') = \mathbb{1}(b \neq b')$。在这种情况下，$R(L)$ 对所有匹配位置中有 L 个字母不匹配的比对求和，也就是说，匹配位置之间的汉明距离为 L。因此，例10中描述的核可以计算为

$$\tilde{k}(X,Y) = \sum_{L=0}^{|X|\wedge|Y|} (C+L)^{-\beta} R(L).$$

下一组 $k_s(b, b') = \delta_b(b')$ 以及 $\ell(b, b') = 1$，因此 $R(L)$ 对所有满足存在 L 个匹配位置且这些位置均具有匹配字母的比对求和。回顾例11中 $\tilde{u}_V$ 的定义。对于序列 V，X，Y，$\tilde{u}_V(X)\tilde{u}_V(Y)$ 是对 X 和 Y 的所有比对求和，其中匹配位置在 X 和 Y 中均为 V，因此，$R(L) = \sum_{|V|=L} \tilde{u}_V(X)\tilde{u}_V(Y)$。因此，例11中描述的核可以计算为

$$\tilde{k}(X,Y) = \sum_{L=0}^{|X|\wedge|Y|} \left(C + \frac{1}{2}(|X|+|Y|) - L\right)^{-\beta} R(L).$$

L. 嵌入核的灵活性

命题47。 （证明命题 *29* 和 *31*）假设 $k(z,z') = \Psi(z-z')$ 是一个平移不变核，并且 Ψ 是定义在 $\mathbb{R}^D$ 上的正连续函数，具有

具有严格正的傅里叶变换。当且仅当 F 是单射时，嵌入核 k 是特征性的且为 C_0-通用。此外，当且仅当 k 具有离散质量时， k 度量化 $\mathcal{P}(S)$ ；而这当且仅当 $F(S)$ 没有聚点时成立，也就是说，不存在序列 $X \in S$，使得 $F(X)$ 属于 $F(S \setminus \{X\})$ 的闭包。

证明 首先我们证明，只有当 F 是单射且 $F(S)$ 没有聚点时， k 才度量化 $\mathcal{P}(S)$。设 $F(X_n) \to F(X)$ 是一列点，且对所有 n 都有 $X_n \neq X$。$k(X_n, X_n) = \Psi(0) = k(X, X)$，并且 $k(X_n, Y) = \Psi(F(X_n) - F(Y)) \to \Psi(F(X) - F(Y)) = k(X, Y)$ 对所有 Y 都成立。因此

$$\left\| \int d\delta_{X_n}(Y) k_Y - \int d\delta_X(Y) k_Y \right\|_k^2 = \|k_{X_n} - k_X\|_k^2 = 2k(X, X) - 2k(X_n, X) \to 0$$

因此， k 并不能度量化 $\mathcal{P}(S)$，因为 $\delta_{X_n} \not\to \delta_X$。

现在我们证明，如果 F 是单射且 $F(S)$ 没有聚点，那么 k 具有离散质量。设 k 不具有离散质量，因此，根据命题 11， $\sum_{m=1}^{M_n} \alpha_{n,\,m} k_{X_{n,\,m}} \to k_X$ 属于 $\mathcal{H}_k$，其中 $X_{n,\,m} \neq X$。记 $\mu_n = \sum_{m=1}^{M_n} \alpha_{n,\,m} \delta_{F(X_{n,m})} - \delta_{F(X)}$。定义 $\hat{}\Psi$ 和 $\hat{}\mu_n$ 分别为 Ψ 和 μ_n 的傅里叶变换，按照 Sriperumbudur 等人（2011）的论证，我们有

$$\int d\xi |\hat{\mu}_n(\xi)|^2 \hat{\Psi}(\xi) = \int \int d\mu_n(x) d\mu_n(y) \Psi(x - y) = \left\| \int d\mu_n(X) k_X \right\|_k \to 0.$$

因此存在一个子序列 $(\hat{}\mu_{n_k})_k$ ，使得 $\hat{}\mu_{n_k} \to 0$ 相对于测度 $\Psi(\xi)d\xi$几乎处处逐点成立。由于 Ψ 严格为正，$\hat{}\mu_{n_k} \to 0$ 相对于勒贝格测度几乎处处逐点成立。这意味着 $\mu_{n_k} \to 0$ 依分布收敛。只有当存在一个序列 $F(X_{n_k,\,m_{n_k}}) \to F(X)$ 时，才可能出现这种情况；也就是说，$F(S)$ 有一个聚点。

最后，我们证明 k 是特征性的且通用，当且仅当 F 是单射。首先，如果 F 不是单射，那么 k就不是特征性的，因为这意味着存在两个具有相同核嵌入的序列。现在假设 F 是单射，且 ν 是一个有限符号测度，并且$\int k_X d\nu(X) = 0$。对所有集合 $C \subset \mathbb{R}^D$，定义 $\mu(C) = \nu(F^{-1}(C))$。那么，根据上面的推理，$\int d\xi |\hat{}\mu(\xi)|^2 \hat{}\Psi(\xi) = \| \int k_X d\nu(X) \|_k = 0$. 因此，$\hat{}\mu = 0$，所以 $\mu = 0$。由于 F是单射，这意味着 $\nu = 0$。因此，对于所有符号测度，映射 $\nu \mapsto \int d\nu(X) k_X$ 都是单射。这等价于 Sriperumbudur 等人（2011）的命题 2 所述的 C_0-全域性。这也意味着该映射对所有概率分布都是单射，因此 k 是特征性的。 ■

M. 稀疏表示学习可能产生缺乏灵活性的嵌入核

考虑使用带有促进稀疏性的正则化的矩阵分解模型来学习序列表示。例如，考虑以下目标函数，

$$\min_{W,B,Z} \sum_{i=1}^{N} \|X_i - \sum_d W_d Z_{id} - B\|_2^2 + \lambda \sum_{d=1}^{D} \sum_{l=0}^{\infty} \sum_{b \in \mathcal{B}} |W_{dlb}| + \beta \sum_{i=1}^{D} \|Z_i\|_2^2. \tag{18}$$

这里，与逐位置比较核（第7节）一样，我们为每个序列 $X \in S$ 添加无限长的终止符号尾部。我们将序列视为独热编码。向量 $Z_i \in$

$\mathbb{R}^D$ 是训练数据的表示，并通过斜率为 W、偏移量为 B 的线性映射映射到独热编码序列空间。嵌入 Z 上采用权重为 β 的 ℓ^2 正则化，W 上采用权重为 λ 的 ℓ^1 正则化。对 W 进行这种套索正则化，会促使其各个条目趋于零，从而使表示更易于解释：Z 的每个维度只会影响少数几个序列位置。模型训练完成后，我们可以用它定义嵌入函数 $F(X) = \text{argmin}_{Z\in\mathbb{R}^D} \|X - \sum_d \hat{W}_d Z_d - \hat{B}\|_2^2 + \beta\|Z\|_2^2$，其中 $\hat{W}$ 和 $\hat{B}$ 是根据训练数据估计出的参数值。

在有限训练数据集上，对于很可能存在位置 $l \in \{1, \ldots, D\}$，我们都有 $W_{dlb} = \hat{W}_{dlb'} = 0$。在这种情况下，一个包含 $X_{lb} = 1$ 的序列 X 将映射为与另一个序列 X' 相同的表征；除了位置 l 上的字母不同（即 $X_{lb'} = 1$）外，该序列与 X 完全相同。换句话说，对于 $X' \neq X$，$F(X) = F(X')$，因此 F 不是单射，而基于 F 的嵌入核既非通用也非特征性的。

N. 缩放随机嵌入

命题48. （命题*33*的证明）考虑一个初始嵌入 F，其中每个$\tilde{F}(X)$对 $X \in S$均从球面上的均匀分布中抽取，$\{x \in \mathbb{R}^D \mid \|x\| \leq 1\}$。那么，对于 $\epsilon > 0$，使用缩放嵌入 $F(X) = |\mathcal{B}|^{(1+\epsilon)|X|/D}\tilde{F}(X)$的核几乎必然度量化$\mathcal{P}(S)$。

证明 按照命题30证明中的相同思路，F是单射的概率为 1。现在，对于任意 N，我们将证明 $\{X \mid \|F(X)\| \leq N\}$ 以概率 1 为有限值。这意味着，对于任意由互不相同的序列构成的序列 $X_1, X_2, \ldots$，当 $n \to \infty$ 时，$\|F(X_n)\|_\infty \to \infty$ 必然成立；特别是，嵌入 $F(S)$ 没有聚点。现在，

$$p(\|F(X)\| \leq N) = p(\|\tilde{F}(X)\| \leq N/|\mathcal{B}|^{(1+\epsilon)|X|/D}).$$

如果 $|\mathcal{B}|^{-(1+\epsilon)|X|/D}N \geq 1$，那么该量为 1。否则，令 V 表示勒贝格测度，令 $B(1)$ 表示单位球 $\{x \in \mathbb{R}^D \mid \|x\| \leq 1\}$，则

$$p(\|F(X)\| \leq N) = \frac{V(|\mathcal{B}|^{-(1+\epsilon)|X|/D}NB(1))}{V(B(1))} = |\mathcal{B}|^{-(1+\epsilon)|X|}N^D.$$

因此，

$$\sum_{X\in S} p(\|F(X)\| \leq N) \leq N^D \sum_{X\in S} |\mathcal{B}|^{-(1+\epsilon)|X|} = N^D \sum_{L=0}^{\infty} |\mathcal{B}|^{-L\epsilon} < \infty.$$

所以到 t 博雷尔-坎特利引理表明，以概率 1，$F(S)$ 没有聚点 . ■

参考文献

10x Genomics。一种探索免疫的新方法：将高通量抗原识别与免疫库和表型相联系。*10x Genomics*，2022年。

Ethan C Alley、Grigory Khimulya、Surojit Biswas、Mohammed AlQuraishi 和 George M Church。利用基于序列的深度表征学习实现统一的理性蛋白质工程。《自然·方法》，16(12):1315–1322，2019年12月。Alan N Amin、Eli N Weinstein 和 Debora S Marks。用于全基因组的生成式非参数贝叶斯模型。神经信息处理系统进展第*34*卷（*NeurIPS 2021*），2021年12月。Miguel A Arcones 和 Evarist Gine。关于 U 统计量和 V 统计量的自助法。《统计年鉴》，20(2):655–674，1992年。Francis Bach、Simon Lacoste-Julien 和 Guillaume Obozinski。集群算法与条件梯度算法的等价性。载于第*29*届国际机器学习大会论文集，ICML'12，第1355–1362页，美国威斯康星州麦迪逊，2012年6月。Omnipress。Balandat、Karrer、Jiang、Daulton、Letham、Wilson 和 Bakshy。BoTorch：一种用于高效蒙特卡洛贝叶斯优化的框架。《神经信息处理系统进展》，2020年。

Luis A Barrera、Anastasia Vedenko、Jesse V Kurland、Julia M Rogers、Stephen S Gissel- brecht、Elizabeth J Rossin、Jaie Woodard、Luca Mariani、Kian Hong Kock、Sachi Inukai、Trevor Siggers、Leila Shokri、Raluca Gordˆan、Nidhi Sahni、Chris Cotsapas、Tong Hao、Song Yi、Manolis Kellis、Mark J Daly、Marc Vidal、David E Hill 和 Martha L Bu- lyk。人类转录因子变异情况调查揭示了普遍存在的DNA结合变化。《科学》，351(6280):1450–1454，2016年3月。

Surojit Biswas、Grigory Khimulya、Ethan C Alley、Kevin M Esvelt 和 George M Church。利用数据高效深度学习进行低样本量蛋白质工程。《自然·方法》，18(4):389–396，2021年。Yutian Chen、Max Welling 和 Alex Smola。通过核集群算法生成超样本。载于第二十六届人工智能不确定性会议论文集，UAI'10，第109–116页，美国弗吉尼亚州阿灵顿，2010年7月。AUAI出版社。Andreas Christmann 和 Ingo Steinwart。非标准输入空间上的通用核。载于 J Lafferty、C Williams、J Shawe-Taylor、R Zemel 和 A Culotta 编，《style id='5'>神经信息处理系统进展，第23卷。Curran Associates, Inc.，2010年。

Bo Dai、Hanjun Dai、Arthur Gretton、Le Song、Dale Schuurmans 和 Niao He。通过双重对偶嵌入进行核指数族估计。载于 Kamalika Chaudhuri 和 Masashi Sugiyama（编），第二十二届国际人工智能与统计会议论文集，第 89 卷，机器学习研究论文集，第 2321–2330 页。PMLR，2019年。

Nicki Skafte Detlefsen、Søren Hauberg 和 Wouter Boomsma。学习蛋白质序列的有意义表示。《自然·通讯》，13(1):1914，2022年4月。

Richard Durbin、Sean R Eddy、Anders Krogh 和 Graeme Mitchison。《生物序列分析：蛋白质与核酸的概率模型》。剑桥大学出版社，1998年4月。

David Duvenaud、James Robert Lloyd、Roger Grosse、Joshua B Tenenbaum 和 Zoubin Ghahramani。通过组合核搜索进行非参数回归中的结构发现。载于第*30*届国际机器学习会议论文集——第*28*卷，ICML'13，第 III–1166–III–1174 页。JMLR.org，2013年6月。

Kenji Fukumizu、Arthur Gretton、Xiaohai Sun 和 Bernhard Sch¨olkopf。条件依赖的核度量。载于神经信息处理系统（*NeurIPS*），2007年。

Kenji Fukumizu、Arthur Gretton、Bernhard Sch¨olkopf 和 Bharath K Sriperumbudur。群和半群上的特征核。《神经信息处理系统进展》，21，2008年。

Futoshi Futami、Zhenghang Cui、Issei Sato 和 Masashi Sugiyama。通过贪心粒子优化进行贝叶斯后验近似。*AAAI*，33(01):3606–3613，2019年7月。

Thomas G¨artner、Peter Flach 和 Stefan Wrobel。关于图核：困难性结果与高效替代方案。载于学习理论与核机器，第129–143页。Springer Berlin Heidelberg，2003年。

Jackson Gorham 和 Lester Mackey。使用核衡量样本质量。载于 Doina Precup 和 Yee Whye Teh 编，第*34*届国际机器学习大会论文集，第70卷，机器学习研究论文集，第1292–1301页。PMLR，2017年。

Arthur Gretton、Kenji Fukumizu、Choon Teo、Le Song、Bernhard Sch¨olkopf 和 Alex Smola。独立性的核统计检验。《神经信息处理系统进展》，20，2007年。

Arthur Gretton、Karsten M Borgwardt、Malte J Rasch、Bernhard Sch¨olkopf 和 Alexander Smola。核双样本检验。机器学习研究期刊，13:723–773，2012年。

David Haussler。离散结构上的卷积核，UCSC CRL。1999。

Ferenc Husz´ar 和 David Duvenaud。最优加权牧群算法是贝叶斯求积。载于第二十八届人工智能不确定性会议论文集，UAI'12，第 377–386 页，美国弗吉尼亚州阿灵顿，2012年8月。AUAI出版社。

T Jaakkola、MDiekhans 和 D Haussler。用于检测远缘蛋白质同源性的判别框架。*J. Comput. Biol.*，7(1-2)：95–114，2000年2月。

Arthur Jacot、Franck Gabriel 和 Cl´ement Hongler。神经切线核：神经网络中的收敛与泛化。载于第*32*届神经信息处理系统国际会议论文集，NIPS'18，第8580–8589页，纽约州雷德胡克，美国，2018年12月。Curran Associates Inc.

P Jorgensen 和 F Tian。《离散再生核希尔伯特空间：采样与狄拉克质量的分布》。机器学习研究期刊，2015年。

Nils M Kriege、Fredrik D Johansson 和 Christopher Morris。图核综述。应用网络科学，5(1):1–42，2020年1月。Rui Kuang、Eugene Ie、Ke Wang、Kai Wang、Mahira Siddiqi、Yoav Freund 和 Christina Leslie。基于谱系的字符串核，用于远缘同源检测和基序提取。*IEEE*计算系统生物学会议论文集，第152–160页，2004年。Jaehoon Lee、Yasaman Bahri、Roman Novak、S Schoenholz、Jeffrey Pennington 和 Jascha Narain Sohl-Dickstein。深度神经网络作为高斯过程。国际学习表征会议，2017年。Christina Leslie、Eleazar Eskin 和 William Stafford Noble。谱核：用于SVM蛋白质分类的字符串核。太平洋生物计算研讨会，第564–575页，2002年。Christina S Leslie、Eleazar Eskin、Adiel Cohen、Jason Weston 和 William Stafford Noble。用于判别式蛋白质分类的失配字符串核。*Bioinformatics*，20(4): 467–476，2004年3月。Chun-Liang Li、Wei-Cheng Chang、Yu Cheng、Yiming Yang 和 Barn´as P´oczos。MMD GAN：深入理解矩匹配网络。载于第*31*届神经信息处理系统国际会议论文集，NIPS'17，第2200–2210页，美国纽约州雷德胡克，2017年12月。Curran Associates Inc. Takuo Matsubara、Jeremias Knoblauch、Fran˛cois-Xavier Briol 和 Chris J Oates。针对似然函数难以处理情形的稳健广义贝叶斯推断。英国皇家统计学会杂志*B*辑：统计方法学，84(3):997–1022，2022年7月。A G D Matthews、M Rowland、Jiri Hron、Richard E Turner 和 Zoubin Ghahramani。宽层深度神经网络中的高斯过程行为。国际学习表征会议，2018年。Radford M Neal。无限网络的先验。载于 Radford M Neal 编，神经网络的贝叶斯学习，第29–53页。Springer New York，纽约州纽约市，1996年。Pascal Notin、Jos´e Miguel Hern´andez-Lobato 和 Y Gal。利用解码器不确定性改进VAE潜在空间中的黑箱优化。《神经信息处理系统进展》，2021年。Mijung Park、Wittawat Jitkrittum 和 Dino Sejdinovic。K2-ABC：基于核嵌入的近似贝叶斯计算。载于 Arthur Gretton 和 Christian C Robert 编，第*19*届国际人工智能与统计会议论文集，第51卷，机器学习研究论文集，第398–407页，加的斯，西班牙，2016年。PMLR。Luc Pronzato。贪心算法最小化最大均值差异的性能分析。*arXiv [stat.ML]*，2021年1月。Luc Pronzato 和 Anatoly Zhigljavsky。贝叶斯求积、能量最小化与空间填充设计。*SIAM/ASA*不确定性量化期刊，8(3):959–1011，2020年1月。

Roshan Rao、Nicholas Bhattacharya、Neil Thomas、Yan Duan、Xi Chen、John Canny、Pieter Abbeel 和 Yun S Song。使用 TAPE 评估蛋白质迁移学习。载于神经信息处理系统进展，第 32 卷，第 9689–9701 页，2019年。Hiroto Saigo、Jean-Philippe Vert、Nobuhisa Ueda 和 Tatsuya Akutsu。使用字符串比对核进行蛋白质同源性检测。*Bioinformatics*，20(11):1682–1689，2004年7月。Gabriele Schweikert、Gunnar R¨atsch、Christian Widmer 和 Bernhard Sch¨olkopf。基因组序列分析中定义域适应算法的实证分析。《神经信息处理系统进展》，第 21 卷，2008年。Louis Shapiro、Renzo Sprugnoli、Paul Barry、Gi-Sang Cheon、Tian-Xiao He、Donatella Merlini 和 Weiping Wang。*Riordan* 群及其应用。Springer International Publishing，2022年。Louis W Shapiro、Seyoum Getu、Wen-Jin Woan 和 Leon C Woodson。Riordan 群。离散应用数学，34(1-3):229–239，1991年11月。

James Benjamin Simon、Sajant Anand 和 Mike Deweese。神经切线核的逆向工程。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），第*39*届国际机器学习大会论文集，机器学习研究论文集第162卷，第20215–20231页。PMLR，2022年。

Rahul Singh、Liyuan Xu 和 Arthur Gretton。多阶段因果推断的核方法：中介分析与动态处理效应。2021年11月。

S¨oren Sonnenburg、Gunnar R¨atsch 和 Konrad Rieck。使用字符串核进行大规模学习。载于大规模核机器。麻省理工学院出版社，2007年。

Bharath K Sriperumbudur、Kenji Fukumizu、Arthur Gretton、Gert R G Lanckriet 和 BernhardSch¨olkopf。再生核希尔伯特空间中概率分布嵌入的核选择与可分类性。神经信息处理系统进展*22*——*2009*年会议论文集，第1750–1758页，2009年。

Bharath K Sriperumbudur、Arthur Gretton、Kenji Fukumizu、Bernhard Sch¨olkopf 和 Gert R G Lanckriet。概率测度的希尔伯特空间嵌入与度量。机器学习研究期刊，11(50):1517–1561，2010年。Bharath K Sriperumbudur、Kenji Fukumizu 和 Gert R G Lanckriet。全域性、特征核与测度的再生核希尔伯特空间嵌入。机器学习研究期刊，12(70):2389–2410，2011年。Samuel Stanton、Wesley Maddox、Nate Gruver、Phillip Maffettone、Emily Delaney、Peyton Greenside 和 Andrew Gordon Wilson。利用去噪自编码器加速生物序列设计的贝叶斯优化。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），第*39*届国际机器学习会议论文集，机器学习研究论文集第162卷，第20459–20478页。PMLR，2022年。

Nora C Toussaint、Christian Widmer、Oliver Kohlbacher 和 Gunnar R¨atsch。在字符串核中利用物理化学性质。*BMC Bioinformatics*，11 Suppl 8:S7，2010年10月。Koji Tsuda、Taishin Kin 和 Kiyoshi Asai。用于生物序列的边缘化核。*Bioinformatics*，18 Suppl 1:S268–75，2002年。Jean-Philippe Vert、Hiroto Saigo 和 Tatsuya Akutsu。用于生物序列的局部比对核。载于计算生物学中的核方法，第131–153页。麻省理工学院出版社，2004年。Eli N Weinstein。生物序列的生成统计方法。博士论文，哈佛大学，安娜堡，美国，2022年。Eli N Weinstein 和 Debora S Marks。用于生成式生物序列预测与预测的结构化观测分布。载于第*38*届国际机器学习会议论文集，第139卷，第11068–11079页。PMLR，2021年。Eli N Weinstein 和 Jeffrey W Miller。贝叶斯数据选择。机器学习研究期刊，24 (23):1–72，2023年。Jason Weston、Bernhard Sch¨olkopf、Eleazar Eskin、Christina Leslie 和 William Stafford Noble。处理核矩阵中的大对角元素。统计数学年刊，55(2):391–408，2003年6月。Christopher Williams。利用无限网络进行计算。《神经信息处理系统进展》，9，1996年。Andrew Gordon Wilson、Zhiting Hu、Ruslan Salakhutdinov 和 Eric P Xing。深度核学习。载于 Arthur Gretton 和 Christian C Robert（编），第*19*届人工智能与统计国际会议论文集，机器学习研究论文集第51卷，第370–378页，加的斯，西班牙，2016年。PMLR。Kevin K Yang、Zachary Wu、Claire N Bedbrook 和 Frances H Arnold。用于机器学习的学习型蛋白质嵌入。*Bioinformatics*，34(15):2642–2648，2018年8月。