混合量子编码：结合振幅和基数编码，增强量子计算中的数据存储和处理能力

巴塔拉普罗巴哈塔姆
信息技术与数字创新学院泰国曼谷北蒙库国王科技大学
s6007011966039@email.kmutnb.ac.th

苏查-斯曼查特
信息技术与数字创新学院泰国曼谷北蒙库国王科技大学
sucha.smanchat@itd.kmutnb.ac.th

2023 第 20 届国际计算机科学与软件工程联合会议（JCSSE） | 979-8-3503-0050-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/JCSSE58229.2023.10201947

摘要-- 这项研究应用比特-分区混合量子编码方法，在量子系统中高效地存储和处理经典数据。通过结合表示索引的振幅编码和表示数据值的基编码，我们引入了一种新技术，充分利用了两种编码方法的优势。我们描述了混合状态的编码和解码过程，强调了这种方法在数据存储和计算效率方面的潜在优势。此外，我们还探讨了解码过程，解决了与量子测量相关的固有不确定性，并讨论了将这种不确定性最小化的策略。我们的研究结果表明，混合编码可以改善量子信息处理任务，使其成为未来量子计算应用中一项前景广阔的技术。还需要进一步研究，以优化编码和解码过程，并探索这种方法在各种量子算法中的全部潜力。

关键词--混合量子编码、振幅编码、基编码、量子信息处理、数据存储、计算效率、量子计算、量子算法

I. 导言

量子计算利用量子力学的独特特性，有望彻底改变我们处理和计算信息的方式。这项技术的目标是在速度和计算能力方面超越经典计算，即量子至上[1]。对于使用经典计算方法需要大量计算资源或时间才能解决的特定计算问题，量子计算机的性能要优于经典计算机。量子计算机能表现出显著优势的一些重要问题包括整数因式分解、非结构化搜索、优化问题、量子模拟和机器学习。量子计算机利用叠加和纠缠等量子现象实现量子优势，这使它们能够同时执行复杂的计算，并且比经典计算机更高效。

利用量子计算能力的一个重大挑战是如何将各种经典问题调整到量子框架中，从而实现量子资源的高效利用。 这项研究他专注于涉及大规模数据的搜索问题，强调开发将经典系统中的数据转换为量子计算机有效工作的方法，从而加快处理时间。

这项研究的重点是为量子计算转换和优化经典搜索问题，通过利用量子计算在处理和分析大规模数据方面比经典计算机更高效的固有优势，在数据密集型应用中充分发挥量子技术的潜力。主要目标是开发高效的数据准备技术，将大规模非结构化数据串（DNA 数据）转换为适合量子计算机处理的格式，而无需考虑噪声或任何物理实施限制。通过为基于门的量子计算机创建专门的算法和数据准备方法，这项研究旨在证明在大规模字符串搜索问题中显著改进数据准备的潜力。

II. 量子计算机与模式搜索

A.量子计算机

根据计算方法，量子计算机主要有两种类型：基于门的量子计算和绝热量子计算。每种方法都有自己的优势和挑战，为处理和操作量子信息提供了独特的方法。

基于门的量子计算[2]是一系列对量子比特进行计算的量子逻辑门。量子门通过利用量子力学原理（如叠加和纠缠）专门设计的操作来操纵量子比特的状态。基于门的量子算法的一个著名例子是肖尔算法、

绝热量子计算 [3] 是一种根本不同的方法，它依赖于量子退火或绝热量子优化原理。这种方法通过缓慢改变量子系统的哈密顿（Hamiltonian），将量子系统从初始状态逐渐演化到最终状态。这个数学结构代表了系统的总能量。其目标是找到与给定问题的最优解相对应的最低能量状态。与基于门的量子计算不同，绝热量子

计算不需要量子门来操纵量子比特。相反，它依赖于系统哈密顿的连续变换，如果进行得足够慢，结果将是最优的。

这两种类型的量子计算机都利用了三种重要的量子现象：量子干涉、量子并行和量子隧道。

量子干涉[4]是一种能让量子计算机操纵量子比特概率振幅的现象。它通过抛弃不正确的解决方案并强化正确的解决方案，使量子算法能够更高效地执行任务。

量子并行[5]利用叠加原理，使量子计算机能够同时处理多个输入。这意味着量子比特或量子比特可以同时表示 0 和 1，而经典比特只能表示 0 或 1。

量子隧道[6]与绝热量子计算尤其相关，因为它能让系统克服能量障碍，更高效地找到优化问题的全局最小值。这种现象允许量子系统同时探索不同的解决方案，摆脱局部最小值并过渡到更优化的配置。通过结合量子隧道技术，绝热量子计算机可以比经典计算方法更高效地解决复杂的优化问题和组合搜索，成为整个量子计算领域的重要组成部分。

B.DNA 序列数据

DNA 序列包含生物体基因构成的信息，但序列本身并没有确定的格式或组织。对 DNA 数据的分析通常涉及模式识别和统计分析，以确定遗传标记或突变。

图像或视频中的像素数据也被视为非结构化数据。图像或视频帧中的每个像素都包含颜色或亮度信息，但像素的排列没有确定的结构。图像和视频分析涉及计算机视觉等技术

和机器学习来识别数据中的对象或模式。本研究特别关注字符串匹配，以便

字符串数据。同样的方法也可能适用于其他类型的非结构化数据，包括像素数据和视频。不过，要使这种方法适用于其他数据类型，还需要进一步的研究和开发。

C.模式匹配的量子算法

Grover 算法[7]的复杂度为 O(√n)，其中 n 为未排序数据库的大小。特别是在处理大型数据库时，这种二次方速度的提高非常明显。不过，Grover 算法是专门为搜索未排序数据库而设计的。如果数据库已排序，在这种量子框架内的数据。在搜索功能方面，格罗弗算法的设计目的是在未排序的数据库中找到目标项的单次出现。

格罗弗的算法可用于加速基因组测序中的模式匹配，例如针对复杂的 RNA 二级结构。这些结构对于理解 RNA 的功能和调控至关重要，但由于其复杂的折叠模式和相互作用，给传统的模式匹配算法带来了巨大挑战。

D.将经典数据编码为 Qbit

许多量子信息处理应用需要将经典数据编码为量子态。

基数编码是用一个单独的量子比特来表示经典数据的每个比特，其状态|0⟩或|1⟩表示值。例如，经典数据 1010 是
|1⟩|0⟩|1⟩|0⟩.振幅编码[8]用量子态的振幅表示经典数据。角度编码用应用于单个量子比特的一组门的旋转角度来表示经典数据。IQP[9]编码用特定电路结构中的门电路模式表示经典数据。哈密顿演化解析[10]通过将经典数据编码成哈密顿，进而演化出所需的量子态。

III. 编码设计

本节将详细介绍我们的 DNA 数据编码方法。首先，我们将详细介绍 DNA 数据的特点，以及我们如何利用量子计算技术解决编码问题。接下来，我们的架构由经典计算机和量子计算机的接口组成，使我们能够利用两者的优势。Loop-Qbit 编码器是这项研究的主要目标。此外，我们还将使用量子计算模拟器来测试我们的方法，然后再进行实验实施。

A.振幅编码

给定波函数。

$$|\dagger\rangle = \sum_{i=0}^{N-1} \alpha_i |i\rangle |0\rangle^{\otimes m} \quad (1)$$

特定情况下，经典算法（如二进制搜索）的复杂度可比 Grover 算法快 O(log n)。对于大规模数据，Grover 算法需要高效的数据准备和编码。此外，准备代表数据库的量子态和实现甲骨文函数也是一项挑战，因为它需要有效的技术来编码和操作量子态。

其中，$N=2^m$ 是 m 个量子比特的可能状态数，i 是索引为 i 的状态，α 是状态 *i* 对应的振幅。

振幅编码是将数据值编码为波函数 $\dagger_A$ 的振幅。

$$|\dagger_A> = \sum^{N-1} \sqrt{\frac{xi}{\sum_{j=0}^{N-1} x_j}} |i> |0>^{\otimes m} \qquad (2)$$

其中，x 是以 *i 为*索引的数据元素的值

B.混合编码

混合编码是一种同时使用基编码和振幅编码对数据进行编码的方法。每个数据点的x_i 值通过振幅编码被编码到一个量子比特的振幅中。每个数据点的值 *b*（0 或 1）通过基数编码被编码到一个量子比特的基数中。数据

由此产生的量子态是所有基态的叠加，每个基态的振幅由相应的 x 值决定。

给定数据 S = [x, b]，其中 x 是 0 到 N-1 范围内的整数，b 是 0 或 1编码步骤如下：

1. 对数据 S 进行归一化处理：$_{xnorm} = x/(N-1)$，$_{bnorm} = b$

2. 根据 (2) 对 x 进行振幅编码，得出

$$|\dagger_x> = \sqrt{xnorm}|0> + \overline{\sqrt{1 - _{xnorm}}}\ |1> \quad (3)$$

3. 对 b

$$|\dagger_b> = |b> \quad (4)$$

4, 应用两种编码状态的张量乘积，得到混合编码

$$|\dagger_{XB}> = |\dagger_x> \otimes |\dagger_b> \quad (5)$$

要解码混合编码状态 $|\dagger_{XB}>$ A 张量积
(5) 的编码状态 $|\dagger_x>$ 和 $|\dagger_b>$，我们可以将混合编码写成

$$|\dagger_{XB}> = (a\,|0> + b\,|1>) \otimes |\dagger_b> \quad (6)$$

其中 $a = \sqrt{xnorm}$, $b = \overline{\sqrt{1 - _{xnorm}}}$

要解码 b，我们可以测量第二个寄存器（即 $_b>$）的计算基础。结果将是状态 |0> 或 |1>，与 b 的原始值相对应。

当 $|\dagger_b> = |0>$ 时，$_{Stateb} = 0$

当 $|\dagger_b> = |1>$ 时，$_{状态}b = 1$

要解码 x，需要应用反向振幅编码变换，这可以通过受控旋转门来实现。应用反向振幅编码后，我们可以测量计算基础中的第一个寄存器，从而得到 $_{xnorm}$ 的估计值。

重建 x 和 b

$$x = \text{round}\ (_{xnorm} * (N - 1))$$

$$b = _{Stateb}$$

IV. DNA 数据的比特分区编码

混合编码结合了从经典数据到量子位编码的多种编码方法。在混合编码中，数据需要通过映射每个 1 和 0 来设置基本编码部分的状态，我们称这一步为比特分区（BP）。

A.DNA 数据

就 DNA 序列数据而言，这涉及在大量遗传信息中搜索特定模式或序列。在计算方面，DNA 序列数据表示为四个核苷酸串：腺嘌呤 (A)、胞嘧啶 (C)、鸟嘌呤 (G) 和胸腺嘧啶 (T)。这些核苷酸被存储在的特定文件格式来管理基因信息。在这项研究中，我们使用以 FASTA 格式存储的 DNA 数据[11]。如图 1 所示，FASTA 格式是一种广泛使用的基于文本的 DNA 序列表示格式。它由单行描述和序列组成，其中每个字符代表一个核苷酸。描述行以">"符号开头，随后是唯一标识符和附加元数据。FASTA 格式简单易读，是存储和共享基因数据的首选格式。

```
>序列1 生物 X 的 DNA 序列
ATCGATCGATCGATCGATCGATCGATCGATCGGA
TCG ATCG
```

图 1.FASTA 文件示例

B. 位分区混合编码

传统的预处理步骤是将 FASTA DNA 序列转换成适合编码的格式。该算法使用 FASTA DNA 作为测试数据，输入长度为 n 的 DNA 序列，并输出每个核苷酸（A、G、C、T）的比特流分区。在经典预处理过程中，算法将四个列表初始化为全零，代表每个核苷酸的分区。然后，对于 DNA 序列中的每个字符，它都会将相应分区列表中的相应条目设为 1，并将其他三个分区列表中的条目设为 0。伪代码如图 2 所示。

```
输入：长度为 n 的 DNA 序列 S
输出：每个核苷酸的位流分区：partition_A、partition_G、
partition_C、partition_T

将四个长度为 n 的列表 partition_A、partition_G、partition_C 和
partition_T 初始化为全部 0

for i from 1 to n do
  if S[i] is 'A' then
     partition_A[i] ← 1
     partition_G[i] ← 0
     分区_C[i] ← 0
     partition_T[i] ← 0
   否则，如果 S[i] 是
                "G"，则
     partition_A[i] ← 0
     partition_G[i] ← 1
        分区_C[i] ← 0
     partition_T[i] ← 0
   否则，如果 S[i] 是
                "C"，则
     partition_A[i] ← 0
     partition_G[i] ← 0
        分区_C[i] ← 1
     partition_T[i] ← 0
   否则，如果 S[i] 是
                "T"，则
     partition_A[i] ← 0
     partition_G[i] ← 0
        分区_C[i] ← 0
     partition_T[i] ← 1
```

图 2.位分区混合编码的伪码

C. 建筑学

将 BP 混合编码应用于 DNA 数据时，需 要 实施四个独立的系统，每个系统对应于

腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶）中的一种。该算法可概括如下：

1. 为每个核苷酸（A、G、C 和 T）初始化四个量子寄存器，每个寄存器由 8 个量子比特组成（每个核苷酸 2 个量子比特）。
2. 通过将 DNA 序列转换为比特流表示来准备输入数据。对于 DNA 序列中的每个字符，将相应分区列表中的相应条目设置为 1，将其他三个分区列表中的条目设置为 0。
3. 使用混合编码方案将输入数据编码到每个核苷酸的量子寄存器中，将索引号编码到量子比特的振幅中，并根据矩阵第二列的值（0 或 1）设置量子比特状态
4. 根据具体应用需要，对编码数据执行量子操作。
5. 测量由此产生的波函数，以获得所需的数据元素概率分布。

D. 结果

BP 混合编码方案对非结构化数据（如 DNA 数据）采用基数和振幅编码。要编码的数据是一个 *i* x 2 矩阵，其中第一列 *i* 代表索引，第二列代表值。目的是将索引号编码到量子比特的振幅中，同时根据矩阵第二列中的值设置量子比特的状态。在这项研究中，我们使用的是 64x2 矩阵；编码时每个核苷酸使用 2 个量子位。由此产生的波函数可以通过测量获得所需的数据元素概率分布。核苷酸 A 的编码和解码示例如图 3 所示，结果分析如图 4 所示。

给定 A = [0110...10]
sa = [[0, 0], [1, 1], [2, 1], [3, 0]...[62,1],[63,0]]

输出状态向量

数据点 [0，0]：状态向量 = ['0.0+0.0j', '0.0+0.0j', '1.0+0.0j', '0.0+0.0j'].
数据点 [1，1]：状态向量 = ['0.0+0.0j', '0.12598816+0.0j', '0.0+0.0j', '0.99203175+0.0j'].
数据点 [2，1]：状态向量 = ['0.0+0.0j', '0.17817416+0.0j', ' 0 . 0+0.0j', '0.98399897+0.0j'].
数据点 [3，0]：状态向量 = ['0.21821789+0.0j', '0.0+0.0j', ' 0 . 97590007+0.0j', '0.0+0.0j'].
数据点 [62, 1]：状态向量 = ['0.0+0.0j', '0.99203175+0.0j', '0.0+0.0j', '0.12598816+0.0j'].
数据点 [63, 0]：状态向量 = ['1.0+0.0j', '0.0+0.0j', '0.0+0.0j', '0.0+0.0j'].

解码与重构
sa = [[0, 0], [1, 1], [2, 1], [3, 0]...[62,1],[63,0]]

图 3.编码器和解码器结果

图 4 中的结果分析展示了一种结合了振幅编码和基数编码的混合编码技术。我们使用振幅编码表示数据点的索引，而基编码则表示实际数据。具体来说，我们用振幅编码对给定数据集的第一列（指数）进行编码，用基数编码对第二列（数据）进行编码。通过这种方法，我们可以表示和分析经典数据，如 DNA 序列

利用量子计算概念以及 Python 和 Qiskit 等工具，将 AGCT 转换为 1 或 0 的比特串。

图 4.结果分析

编码过程包括对第一列中的索引值进行归一化处理，并将这些归一化值编码为量子态的振幅。然后，我们使用基础编码，以量子比特的二进制表示法来表示数据集的第二列。这样，我们就能以状态矢量和基态的形式来表示数据集，而状态矢量和基态是量子的基本要素。

归一化系数：N = 64

数据点 [0，0]：归一化值： 0/63 数据点 [1，1]：正常化值： 1/63 数据点 [62，1]：正常化值： 62/63 数据点 [63，0]：规范化值： 63/63

数据点 [0, 0]

状态向量：[0.+0.j, 0.+0.j, 1.+0.j, 0.+0.j] 解码第一个量子比特振幅：|0|^2 + |0|^2 = 0

基态和振幅：|00⟩:0, |01⟩:0, |10⟩:1, |11⟩:0

最高振幅为基态 |10⟩；第二个量子位解码为 0。 重构 $_{x0}$ = 0/63 * 63 = 0
重建 $_{b0}$ = 0

数据点 [1, 1]

状态向量：[0.+0.j，0.12598816+0.j，0.+0.j，0.99203175+0.j)

解码第一个量子位振幅：|0|^2 + |0.12598816|^2 ≈ 1/63 基态和振幅：|00⟩:0, |01⟩:0.12598816, |10⟩:0, |11⟩:0.99203175

最高振幅为基态 |11⟩；第二个量子位解码为 1。 重构 $_{x1}$ = 1/63 * 63 = 1
重建 $_{b1}$ = 1

数据点 [2, 1]

状态向量：[0.+0.j，0.17817416+0.j，0.+0.j，0.98399897+0.j)

解码第一个量子位振幅：|0|^2 + |0.17817416|^2 ≈ 2/63 基态和振幅：|00⟩:0, |01⟩:0.17817416, |10⟩:0, |11⟩:0.98399897

最高振幅为基态 |11⟩；第二个量子位解码为 1。 重构 $_{x2}$ = 2/63 * 63 = 2
重建 $_{b2}$ = 1

数据点 [3, 0]

状态向量：[0.21821789+0.j，0.+0.j，0.97590007+0.j，0.+0.j]

解码第一个量子位振幅：|0.21821789|^2 + |0|^2 ≈ 3/63 基态和振幅：|00⟩:0.21821789, |01⟩:0, |10⟩:0.97590007, |11⟩:0

最高振幅为基态 |10⟩；第二个量子位解码为 0。 重构 $_{x3}$ = 3/63 * 63 = 3
重建 $_{b3}$ = 0

数据点 [62, 1]

状态向量：[0.+0.j，0.99203175+0.j，0.+0.j，0.12598816+0.j)

解码第一个量子位振幅：|0|^2 + |0.99203175|^2 ≈ 62/63 基态和振幅：|00⟩:0, |01⟩:0.99203175, |10⟩:0, |11⟩:0.12598816

最高振幅为基态 |01⟩；第二个量子位解码为 1。 重构 $_{x62}$ = 62/63 * 63 = 62
重建 $_{b62}$ = 1

数据点 [63, 0]

状态向量：[1.+0.j, 0.+0.j, 0.+0.j, 0.+0.j] 解码第一个量子比特振幅：|1|^2 + |0|^2 = 1

基态和振幅：|00⟩:1, |01⟩:0, |10⟩:0, |11⟩:0

最高振幅为基态 |00⟩；第二个量子位解码为 0。 重构 $_{x63}$ = 1 * 63 = 63
重建 $_{b63}$ = 0

sa = [[0，0]，[1，1]，[2，1]，[3，0]...[62，1]，[63，0]]

计算。利用这种技术，我们可以保持量子计算机所利用的量子现象：量子干涉、量子并行和量子隧道。

混合编码结合了表示数据点索引的振幅编码和表示实际数据的基数编码。它平衡了每种方法的优缺点。振幅编码能有效地表示连续数据，但可能存在分辨率低、易受噪声影响等问题。另一方面，基数编码能更稳健地表示离散数据，但需要更多的量子比特来处理高维数据。通过对数据点的索引使用振幅编码，对实际数据使用基数编码，我们可以利用振幅编码对连续索引值的高效性，同时利用基数编码对离散数据的鲁棒性。这种组合减轻了单独使用其中一种编码方法的一些缺点，为在量子系统中表示经典数据提供了一种更通用、更实用的方法。

E. 结论

对字符串使用已实现的 BP 混合编码的优势在于，它允许我们将字符串中的每个字符表示为一个单独的数组，可用于排序和搜索等特定操作。例如，如果我们有一个庞大的字符串数据集，而我们又想搜索这些字符串中出现的所有特定字符，那么 BP 混合编码就能提高这项任务的效率。这种编码还有助于减少字符串数据的存储需求。在某些情况下，字符串可能包含重复的字符或单词模式。BP 混合编码可以识别这些模式并将其存储为单独的数组，然后可以在多个字符串中重复使用，从而大大节省内存。这种编码方法可以与量子算法集成，因此我们可以充分利用量子技术的潜力，提高大规模数据上各种操作的性能。

总之，BP 混合编码是一种灵活高效的字符串数据表示方式，有助于提高对该数据进行各种操作的性能，同时降低存储要求。

F. 今后的工作

我们未来的工作目标是扩展 BP 混合编码方法的功能，以更好地处理各种类型的数据、包括带有色彩深度值的像素数据。它的改进将使混合编码方法更具通用性，适用于各种应用。例如，像素数据可以使用颜色量化技术来减少图像中使用的颜色数量。

BP 混合编码可通过基于门的量子计算机和基于退火的量子计算机实现。门式量子计算机具有很高的灵活性，可以实现任意的单元操作，因此适合许多量子算法。另一方面，基于退火的量子计算机具有更简单的架构，更适合优化问题。

参考资料

[1] C.Palacios-Berraquero, L. Mueck, and D. M. Persaud, "Instead of 'supremacy' use 'quantum advantage'," *Nature,* v o l . 576, no. 7786, pp.

[2] K.K.Michielsen、M.Nocon、D.Willsch、F.Jin、T.Lippert 和 H.De Raedt, "Benchmarking gate-based quantum computers," *Computer Physics Communications, vol.* 220, pp.

[3] C.C. C. McGeoch、R. Harris、S. P. Reinhardt 和 P. I. Bunyk，《基于退火的实用量子计算》，《*计算机*》，第 52 卷，第 6 期，第 38-46 页，2019 年 6 月。doi:10.1109/MC.2019.2908836

[4] R.R. C. Liu、B. Odom、Y. Yamamoto 和 S. Tarucha，"电子碰撞中的量子干涉"，《*自然》，第* 391 卷，第 6664 期，第 263-265 页，1998 年 1 月。

[5] P.W. Shor，"量子计算机上*的*质因式分解和离散对数的多项式时间算法"，《*SIAM 计算学报》，第* 26 卷，第 5 期，第 1484-1509 页，1997 年 10 月。5, pp.

[6] M.Razavy, *隧道的量子理论*。新泽西州河边：世界科学出版社，2003 年。

[7] L.K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing - STOC '96*, Philadelphia, Pennsylvania, United States：ACM Press, 1996, pp.

[8] M.M. A. Nielsen 和 I. L. Chuang，《*量子计算与量子信息*》，10 周年纪念版。剑桥；纽约：剑桥大学出版社，2010 年。

[9] D.Beaulieu, D. Miracle, A. Pham, and W. Scherr, "Quantum Kernel for Image Classification of Real World Manufacturing Defects." arXiv, December 2022.[Online] http://arxiv.org/abs/2212.08693.

[10] S. Stanisic 等：《利用量子计算机上的可扩展算法观测费米-哈伯德模型的基态特性》，《*自然-通讯》第* 13 卷第 1 期第 5743 页，2022 年 10 月。

[11] D. Pratas、M. Hosseini 和 A. J. Pinho，"Cryfa：A Tool to Compact and Encrypt FASTA Files," in *11th International Conference on Practical Applications of Computational Biology & Bioinformatics,* F. Fdez- Riverola, M. S. Mohamad, M. Rocha, J. F. De Paz, and T. Pinto, Eds., Cham：施普林格国际出版公司，2017 年，第 305-312 页。