

基于希尔伯特分形的基因组序列压缩算法

陈 旻¹, 王开云², 吴建国³, 李建军³

(1. 云南大学 信息学院, 云南 昆明 650061; 2. 昆明学院 学报编辑部, 云南 昆明 650214;
3. 云南警官学院 信息网络安全学院, 云南 昆明 650223)

摘要: 给出一种基于希尔伯特分形的基因组序列压缩算法. 为充分利用碱基间的相关性, 算法首先使用希尔伯特分形曲线将基因组序列从一维映射到二维, 从而得到映射图像. 再对映射图像使用 Context 加权建模熵编码技术进行压缩. 在 Context 加权中, 权值的确定与各 Context 模型对应的描述长度有关. 当接收端收到压缩图像后, 对其进行解码, 然后根据拟希尔伯特逆矩阵将映射图像转为一维, 从而获得基因组序列. 实验结果表明, 尽管基于希尔伯特空间填充的二维基因组 Context 建模会引入无效编码区, 但最终的压缩结果要略好于其他直接进行 Context 建模的算法.

关键词: 基因组压缩; 希尔伯特空间填充; Context 加权; 描述长度

中图分类号: TP919.1 **文献标识码:** A **文章编号:** 1674-5639(2014)06-0042-05

The Genome Sequence Compression Algorithm Based on the Hilbert Grouping

CHEN Min¹, WANG Kai-yun², WU Jian-guo³, LI Jian-jun³

(1. Information College, Yunnan University, Yunnan Kunming 650061, China;

2. Editorial Department of Journal, Kunming University, Yunnan Kunming 650214, China;

3. Information Security College, Yunnan Police Officer Academy, Yunnan Kunming 650223, China)

Abstract: The genome compression algorithm based on the Hilbert grouping is proposed to fully utilize the correlations among the basic groups. The Hilbert grouping curve is first used in algorithm to map the genome sequence from one dimension into a new 2-D to obtain the image then compressed by the Context weighting modeling encode technology. In Context weighting, the values of weights are decided by the corresponding description length of the Context models. When the receiver obtains the compressed image and decoded, the supposed Hilbert inverse matrix is used to turn the mapping image into one dimension so as to get genome sequence. The experiments results indicate that although the valid coding area will be led by 2-D genome sequence Context modeling based on the Hilbert space filling, the final compression results by our algorithm are a bit better than other results by the direct Context modeling algorithm

Key words: genome sequence compression; Hilbert space filling; context weighting; description length

基因组序列压缩能够降低基因组数据存储代价, 从而提高基因组存储效率. 尽管基于字典压缩的一大类替换压缩算法^[1-3]能够获得较高的压缩效率, 但随着加入字典的序列越来越多, 最终会让编码每个字典条目(index)的码长增加. 特别地, 在文献[3]中, 为了提高压缩效率, 该算法直接对当前序列与字典中存在序列间的差异碱基进行编码, 从而获得目前为止最高的压缩效率. 但这个结果事实上忽略了字典本身的传递代价. 也就是说, 如果接收端没有相应字典, 则无法解码. 这意味着除了传递基因组序列本身, 该算法还必须对字典编码, 这样将大大降低其压缩效率. 因此, 基于字典类的压缩算法并不能保证较高的压缩效率^[4].

另一方面, 基于 Context 建模的熵编码技术被广泛应用于数字信号压缩. 使用 Context 建模熵编码技术对基因组序列进行压缩, 可以避免对字典编码产

生的代价, 从而改善基因组压缩效果. 该压缩技术本质上是利用碱基间的相关性来获得对待编码碱基统计特性的有效估计, 并借助算术编码器获得较短编码码长. 然而, 由于基因组序列中大量存在插入删除片段(indels), 导致直接采用 Context 建模效果并不好. 文献[5]中, Pihno 指出, 使用有限阶 Context 模型能够获得比高阶 Context 建模更好的压缩结果. 为了充分利用碱基间的相关性而又不至于增加 Context 模型阶数, Context 加权被用于基因组 Context 熵编码压缩. 然而, Context 加权依赖于权值, 不同权值带来的压缩效率不一样. 因此, Context 加权的一项重要工作就是确定权值. 文献[5~7]给出了不同的确定权值的方法. 特别在文献[7]中, 给出了根据 Context 模型描述长度进行权值确定的方法, 将加权与描述长度进行了直接关联, 从而获得较理想的压缩效果. 然而, 当给定 Context 模型和训练序列后,

收稿日期: 2014-11-02

基金项目: 云南省自然科学基金青年基金资助项目(2013FD042); 云南大学研究生重点科研基金资助项目(yuny201383).

作者简介: 陈旻(1982—), 男, 云南昆明人, 工程师, 博士研究生, IEEE 会员, IEICE 会员, 主要从事信息传输理论研究.

模型描述长度即被固定. 这意味着如果参与加权的Context模型描述长度均过长, 则加权后也不可能获得一个描述长度较短的编码模型. 这就限制了Context加权的应用. 其实, 模型描述长度与Context建模方法有关, 而Context建模归根结底是要充分利用碱基间的相关性. 如果相关性越强, 则有可能使得建模后Context模型描述长度最短.

然而, 大量的实验表明, 基因组序列中各碱基间并不满足直接相关. 也就是说, 基因组序列是非平稳信源. 对于此类信源的建模, 要想直接获得其空间相关性是困难的. 但是, 根据信号处理的知识可以知道, 任何的信号降维操作, 都会带来信号相关性的降低. 而增加信号的表述维度, 则有可能发现信号的更多相关信息. 对于基因组序列而言, 可以考虑将其映射到更高维, 从而发现并利用碱基间的相关性. 将一维信号映射到二维的方法较多, 但使用分形曲线, 可以简化映射过程. 尽管分形曲线未必保证映射结果最优, 但本文中, 我们并不依赖最优性, 而是只要让碱基间相关性增强即可.

本文中, 我们使用希尔伯特分形曲线(也称为希尔伯特空间填充曲线)来将基因组序列从一维映射到二维. 对于映射过程, 采用文献[8]给出的方法进行, 首先构建希尔伯特空间填充矩阵, 然后再进行映射升维. 在获得映射图像后, 对该图像进行编码. 对于接收方而言, 收到编码后的图像后, 对其进行解码, 然后按照逆希尔伯特变换矩阵, 再将二维图像映射回一维, 从而获得基因组序列. 同时, 为了尽可能多的利用二维图像的相关信息, 仍然采用加权Context建模对基因组序列进行压缩.

1 最优化 Context 建模

令序列 $s_n = x_n, \dots, x_0$ 表示待编码的碱基序列. 对 s_n 进行熵编码后, 序列的码长 L_n 可以由(1)式给出:

$$L_n = -\log_2 p(x_n, \dots, x_0), \quad (1)$$

其中, $p(x_n, \dots, x_0)$ 表示碱基序列 s_n 的联合概率. 因此, L_n 最小实质上等价于序列的联合概率值最大. 然而, 在实际应用中, 序列 s_n 的联合概率并不知道, 但由信息论可知, 联合概率 $p(x_n, \dots, x_0)$ 满足链式规则, 如(2)式所示:

$$p(x_n, \dots, x_0) = p(x_n | x_{n-1}, \dots, x_0) \times p(x_{n-1}, \dots, x_0), \quad (2)$$

将(2)式两边取对数, 可以得到编码码长的递推式(3):

$$L_n = L(x_n) + L_{n-1}, \quad (3)$$

其中, $L(x_n) = -\log p(x_n | x_{n-1}, \dots, x_0)$ 表示对当前碱基 x_n 进行编码的码长. 对序列 s_n 中的 $n+1$ 个碱基进行编码后, 总共可以得到 n 个与(3)式类似的码长表达式. 将这些递推式进行相加, 可以得到(4)式:

$$L_n - L_0 = -\sum_{i=1}^n \log p(x_i | x_{i-1}, \dots, x_0). \quad (4)$$

欲使 L_n 最小, 其实等价于令每个 $p(x_i | x_{i-1}, \dots, x_0)$ 最大. 这也就意味着条件序列 x_{i-1}, \dots, x_0 要能够提供尽可能多的信息来降低当前碱基 x_i 的不确定性. 然而, 在实际应用中, 条件概率 $p(x_i | x_{i-1}, \dots, x_0)$ 是由统计估计得到的. 随着编码的进行, 条件序列 x_{n-1}, \dots, x_0 的长度将越来越长. 因此, 对每个碱基 x_i 都进行满序列条件概率估计是不现实的. 一种可行的方法是, 使用当前碱基 x_i 的过去 K 个碱基 x_{i-1}, \dots, x_{i-K} 来估计其条件概率 $p(x_i | x_{i-1}, \dots, x_{i-K})$, 并让该条件概率尽可能接近 $p(x_i | x_{i-1}, \dots, x_0)$. 但这样一来, 想要保证每个 $p(x_i | x_{i-1}, \dots, x_{i-K})$ 都最大是不可能的. 通常的做法是转为对概率 $p(x_i | x_{i-1}, \dots, x_{i-K})$ 所在的条件概率分布 $P(x_i | x_{i-1}, \dots, x_{i-K})$ 进行估计, 并寻找某种建模方法, 使得条件概率分布的条件熵最小. 即给定阶数 K , 使得条件熵 $H(x_i | x_{i-1}, \dots, x_{i-K})$ 最小, 同时还要让其趋近于满条件序列时的熵 $H(x_i | x_{i-1}, \dots, x_0)$. 这个要求是严格的, 但对于 K 阶记忆信源来说, 其对应条件熵满足(5)式:

$$H(x_i | x_{i-1}, \dots, x_{i-K}) \approx H(x_i | x_{i-1}, \dots, x_0). \quad (5)$$

换言之, 如果令 K 阶条件熵最小, 则意味着 $H(x_i | x_{i-1}, \dots, x_0)$ 最小. 同时, 根据条件降低熵值理论, 如果有 $j > i$, 则有(6)式:

$$H(x_j | x_{j-1}, \dots, x_0) \leq H(x_i | x_{i-1}, \dots, x_0). \quad (6)$$

这意味着对于 K 阶记忆信源, 如果有 $j > i > K$, 则就算再增加条件位数量也并不能带来条件熵的明显降低. 换言之, 对 K 阶记忆信源进行建模, 关键在于找到合适的 K , 使得条件熵最小. Context建模熵编码技术正是利用上述原理, 通过选取当前碱基 x_i 的过去 K 个碱基 x_{i-1}, \dots, x_{i-K} 来构建条件概率分布 $P(x_i | x_{i-1}, \dots, x_{i-K})$, 并寻找最小化 $H(x_i | x_{i-1}, \dots, x_{i-K})$ 的办法.

采用这样的方法其实是合理的, 因为条件熵 $H(x_i | x_{i-1}, \dots, x_{i-K})$ 最小, 意味着对序列 s_n 中的每个碱基进行编码时, 有可能获得较短的平均码长, 从而提高压缩效率. 基于以上考虑, 则(4)式可以改写为:

$$L_n = -\sum_{i=0}^n \log P(x_i | x_{i-1}, \dots, x_{i-K}). \quad (7)$$

在文献[9]中, L_n 也称为给定 K 阶 Context 模型下序列 s_n 的描述长度, 并具有如(8)式所示的描述形式:

$$L_n = (n+1)H(x_i | x_{i-1}, \dots, x_{i-K}) + \Delta, \quad (8)$$

其中, Δ 表示使用 Context 建模熵编码技术进行编码时引入的模型代价. 而 $(n+1)H(x_i | x_{i-1}, \dots, x_{i-K})$ 代表了使用 Context 模型对序列 s_n 中的 $n+1$ 个碱基进行编码时的理论描述长度, 我们也称之为理想描述长度. 在文献[9]中, 我们对模型代价 Δ 进行了详细讨论, 包括其性质和最小化的方法. 通过降低模型代价, 有可能使得序列的描述长度趋于理想描述长度, 从而获得较好的压缩效果. 事

实上,根据我们之前的研究可以知道,如果序列 s_n 的描述长度 L_n 最短,那么对将来碱基进行编码时,有可能获得较短的编码码长.换言之,对模型进行优化的本质其实是已编码序列的描述长度最短.这样称为最小描述长度模型优化^[10].但从(8)式中不难看出,除了降低模型代价 Δ 以外,如果能够降低理想描述长度 $(n+1)H(x_i|x_{i-1}, \dots, x_{i-k})$,则同样可以降低描述长度 L_n .降低理想描述长度 $(n+1)H(x_i|x_{i-1}, \dots, x_{i-k})$,实际上等价于降低条件熵 $H(x_i|x_{i-1}, \dots, x_{i-k})$,而条件熵又与条件 x_{i-1}, \dots, x_{i-k} 有关.如果 x_{i-1}, \dots, x_{i-k} 与当前碱基 x_i 强相关,则意味着条件能够为 x_i 的估计提供尽可能多的信息,从而降低 x_i 的不确定性.相反,如果条件与 x_i 无关,则由于增加条件带来的条件熵 $H(x_i|x_{i-1}, \dots, x_{i-k})$ 降低不明显.因此,最优 Context 建模可以描述为:选择与当前符号最相关的 K 个过去符号来对当前符号的分布特性进行估计,并使得到的条件概率分布具有最小熵值.

然而,在实际应用中,这样的选取是困难的.尽管对图像一类相关信源,由于图像相邻像素间具有局部相关性,因此,直接选取与当前信源符号邻近的

K 个像素点作为条件确实能够保证条件位与当前符号之间的强相关性.同时,借助文献[10]中给出的动态建模方法,在图像编码中,最优 Context 建模确实能够在一定程度上获得较好的条件概率分布估计,从而降低编码码长.然而,对于非平稳信源(例如基因组)来说,直接选取相邻碱基做条件,并不能保证强相关性的获得.我们在前期实验中发现,直接选取邻近碱基做条件并不能获得较为理想的压缩效果.尽管借助 Context 加权和 Context 量化,压缩性能有所提高,但效果并不明显.

为了提升 Context 建模的有效性,本文提出使用空间填充的办法将基因组序列映射到二维,然后对得到的基因组映射图像进行压缩,即一维向二维的映射,使用希尔伯特空间填充.

2 希尔伯特空间填充

希尔伯特空间填充曲线(Hilbert space-filling curve)能够将一维信号映射到二维.其表述为,通过相应希尔伯特空间填充矩阵,将一维信号填入二维空间相应位置,从而实现升维映射.图1给出了不同阶数下的希尔伯特空间填充曲线.

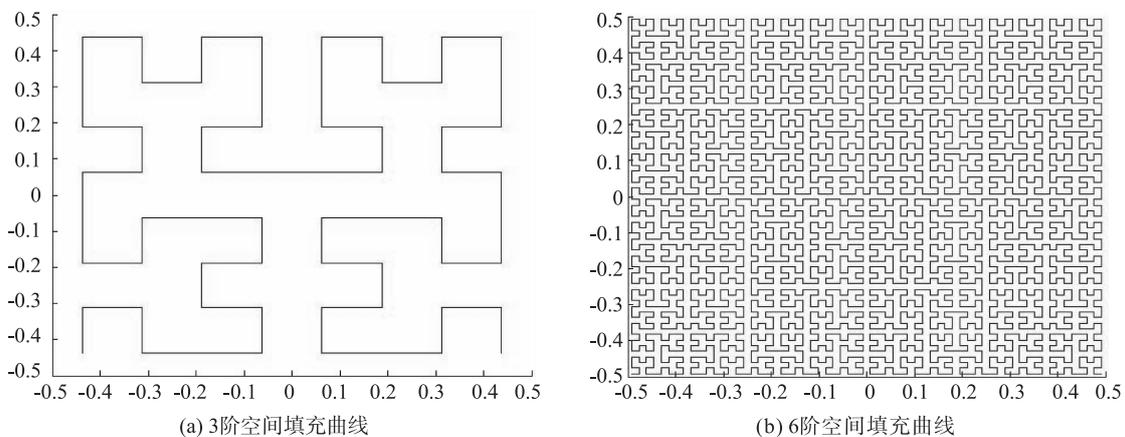


图1 不同阶数的希尔伯特空间填充曲线

从图1中不难看出,随着阶数的升高,空间填充曲线能够覆盖越来越广的空间区域,从而实现信号从一维向二维的映射.

在将基因组序列映射到二维图像的过程中,希尔伯特空间填充曲线其实指示了每个碱基在空间的分布位置.换言之,序列的第 i 个碱基,对应二维空间的位置

为 $l(x, y)$.这种映射关系,可以使用希尔伯特空间填充矩阵来描述.根据文献[8]的研究,希尔伯特空间填充矩阵由(9)式获得.设1阶希尔伯特空间填充矩阵为:

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix},$$

则 $k+1$ 阶空间填充矩阵为:

$$\begin{cases} \begin{bmatrix} H_{2k} & 4^k E_{2k} + H_{2k} \\ 4^{(k+1)} E_{2k} - H_{2k} & (3 \times 4^k + 1) E_{2k} - (H_{2k})^T \end{bmatrix}, k \text{ 为偶数,} \\ \begin{bmatrix} H_{2k} & (4^{k+1} + 1) E_{2k} - H_{2k} \\ 4^k E_{2k} + H_{2k}^T & (3 \times 4^{k+1}) E_{2k} - (H_{2k})^T \end{bmatrix}, k \text{ 为奇数.} \end{cases} \quad (9)$$

其中, E_{2k} 代表维度为 2^k 的单位矩阵.根据(9)式,可以得到任意阶的希尔伯特填充矩阵.2阶希尔伯特

空间填充矩阵如(10)式所示:

$$\begin{bmatrix} 1 & 2 & 15 & 16 \\ 4 & 3 & 14 & 13 \\ 5 & 8 & 9 & 12 \\ 6 & 7 & 10 & 11 \end{bmatrix} \quad (10)$$

空间填充矩阵代表序列中各碱基在空间的位置. 例如矩阵(10)中值为 4 的元素代表着序列中的

第 4 个碱基处于矩阵的第 2 行第 1 列. 换言之, $s_i = l_i(x, y)$. 其中, $l_i(x, y)$ 代表矩阵中值为 i 的元素, 其坐标为 (x, y) .

在获得填充矩阵以及给定序列映射方法后, 可以得到映射后的基因组序列的映射图像, 如图 2 所示.

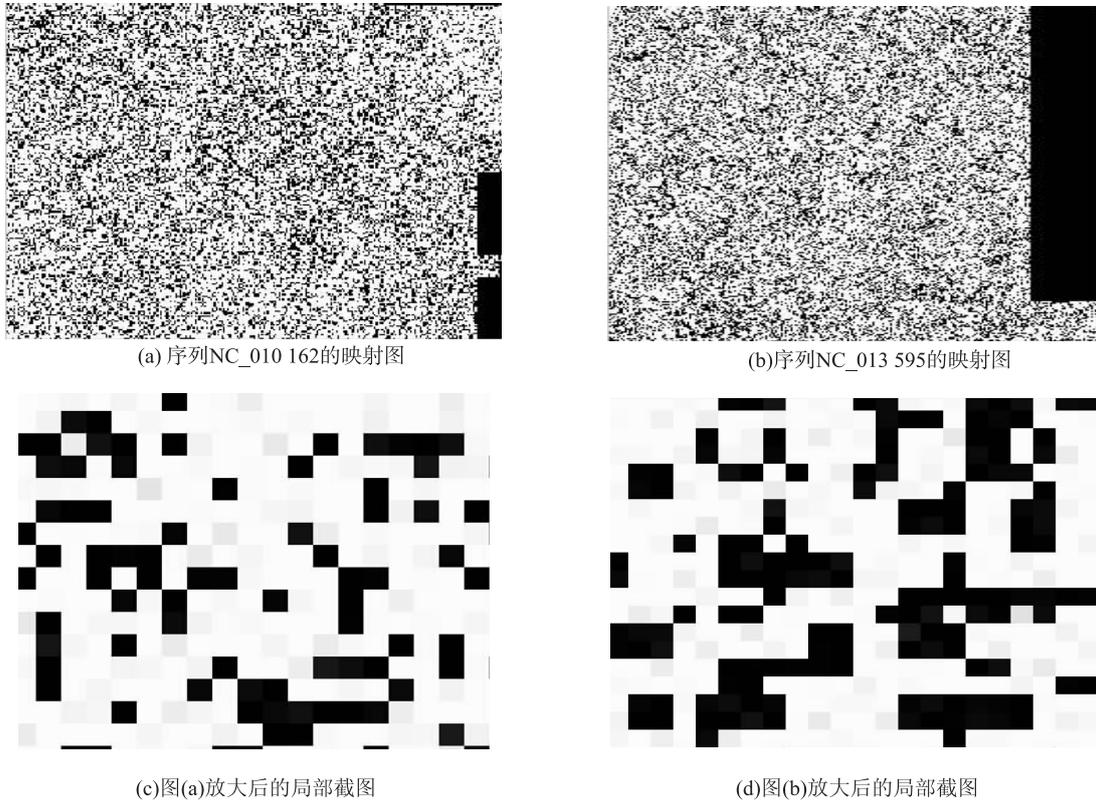


图2 两条基因组序列映射到二维空间后的映射图像

从图 2 可以看出, 通过希尔伯特空间填充, 基因组序列碱基间的相关性得以增强. 特别在图 2(c) 和图 2(d) 中, 可以明显看到 4 种碱基的局部分布. 不难看出, 图 2(c) 和图 2(d) 中, 像素间存在纹理特性, 这意味着碱基间存在较多相关性, 在编码时可以充分利用. 在下节中, 我们将给出基于空间填充算法的基因组序列压缩方法.

3 基于空间填充升维的基因组序列 Context 建模压缩

在获得映射图后, 对基因组序列的压缩转为对映射图的压缩. 为了获得更好的压缩效果, 我们采用文献[9]中提出的加权 Context 建模方式, 对不同方式建模获得的编码模型进行加权. 由于将基因组序列二维化, 可以利用的相关性增强. 而在文献[4]中, Pinho 提出使用有限阶建模将获得更好的压缩效果. 为了保证尽可能使用条件间的相关性又不导致模型阶数的增加, 加权是一个有效的方法.

首先建立 Context 模型. 在本文中, 我们为当前

碱基建立两个 Context 模型(4 阶和 3 阶), 其各自使用的条件碱基模板如图 3 所示, 图 3 中的 x_i 为当前编码碱基.

C_2	C_1	x_i
	C_3	C_4

C_2	C_1
C_3	x_i

(a) 第1个Context模型使用的条件位模板 (b) 第2个Context模型使用的条件位模板

图3 两个Context模型建模时使用的模板

我们采用文献[9]的方法确定权值. 每个 Context 模型对应的权值由(11)式得到:

$$w_i = \frac{L_{i, \max} - L_i}{\sum_{i=1}^2 L_i - L_{i, \max}} \quad (11)$$

其中, L_i 代表第 $i, i = 1, 2$ 个 Context 模型的描述长度, $L_{i, \max}$ 代表两个 Context 模型的描述长度中较大的一个. 在对每个碱基进行编码时, 不需要按照碱基原来在序列中的顺序进行, 而是按照对图像编码时采

用的扫描方式顺序编码(也就是按照行顺序对碱基进行编码)或者按列扫描的方式进行编码.对于解码端,其接收到的其实是基因组映射图像的相应像素(像素值等于实现约定好的碱基编码值0,1,2,3).等到整幅映射图像编码完毕,再按照逆希尔伯特填充将图像重新映射回一维即可.

此外,对于该算法需要考虑两个细节.首先,在使用希尔伯特空间填充矩阵对基因组序列进行升维映射时,并不一定能够保证图像的像素点个数等于基因组序列中碱基的数量.这就意味着在二维图像中,一定存在一部分区域的像素不表示碱基符号,我们称这部分区域为“无效编码区”.但编码时,无效编码区中的像素不能忽略,需要参与编码.这似乎会让最终的码长增加.事实上,如果我们一开始就将图像的全部像素点初始化为某个碱基,在进行二维映射时,如果某个图像像素位置 $l(x,y)$ 上有新的碱基到达,则将该像素点替换为当前碱基.如果 $l(x,y)$ 处于无效编码区,则不替换.这样一来,处于无效编码区中的碱基周围将全是相同的符号,换言之,对其编码需要的码长其实并不会太长.相比起由于相关性增强而带来的码长降低,这部分代价是可以接受的.

其次,使用希尔伯特变化进行升维映射时,需要事先指定空间填充曲线的起点,不同起点得到的空间填充结果不一致.为了保持收发双方具有相同的曲线形式,可以在编码前就约定好起点位置.如此一来,收发双方将不需要面对因扫描曲线不同而带来的无法解码问题.

使用空间填充进行扩展的基因组压缩算法步骤为:

步骤1 确定待编码碱基的总数量,并向接收

端发送这个数值;

步骤2 确定希尔伯特矩阵的最小阶,使得映射后图像的像素总数与碱基长度的差最小;

步骤3 对基因组序列进行升维映射,得到映射图;

步骤4 使用图3所示的条件位置进行Context建模.在对碱基编码时,使用Context加权获得编码模型,从而驱动算术编码器进行编码;

步骤5 待映射图中的每个像素(包含有效编码区和无效编码区)均编码传输完毕,在接收端重构映射图像.并使用逆希尔伯特矩阵,恢复出基因组序列.

4 实验

为验证本文Context建模方法的可行性,我们将其应用于基因组序列压缩.与文献[9]相同的3条DNA序列NC_013 595, NC_013 131和NC_010 162被用作编码序列.同时,另外2条DNA序列NC_013 929和NC_014 318用作训练序列.以上5条序列均可以在National Center for Biotechnology Information (NCBI)^[11]获得.

在实验中,首先按照图3所示的模板分别进行4阶和3阶Context建模.然后将训练序列进行二维映射,并使用映射得到的二维图像对两个Context模型进行训练.随后将编码序列进行希尔伯特二维映射,并对映射图进行编码.在编码过程中,采用Context加权来获得编码模型,权值的确定方法由(11)式给定.3条DNA序列压缩结果如下表1所示.为方便描述,我们在表1中仅以其编码bit率(单位:bit/碱基)作为结果描述.为方便对比,文献[5~7]的压缩结果同样列于表中.

表1 3个DNA序列编码结果对比

序列	序列尺寸	编码结果/(bit · base ⁻¹)			
		文献[5]方法	文献[6]方法	文献[7]方法	本文方法
NC_013 595	10 341 314	1.796	1.759	1.742	1.740
NC_013 131	10 468 872	1.817	1.779	1.770	1.765
NC_010 162	13 033 779	1.755	1.743	1.732	1.725

从表1可以看出,借助本文Context建模方法,基因组序列的压缩效率得以提高.同时,本文以细菌DNA序列为编码序列,该类序列中各碱基间的相关性本来就不强,难以获得较好压缩结果.但通过进行二维映射后,碱基间的相关性得以增强,最终使得编码性能得以提升.此外,为压缩这3条DNA序列,我们至少需要构建12阶的希尔伯特矩阵.而表1中,本文算法的压缩结果包括了无效编码区符号的码长.表1中本文的“bit/碱基”由“ $L_{\text{图像码长}}/\text{碱基数量}$ ”计算.换言之,码长包括了无效编码区符号码长,而计数值则是使用碱基本身的数量.这样一来,本文方法获得的码长将有可能更长.但事实上,由于相关性

增强而带来的编码收益要优于因为引入无效编码区而造成的代价.因此,在表1中,本文最终压缩效果要略好于过去算法.而且,对于序列NC_010 162,由于其碱基数量更接近映射图中像素的总数,因此,相比另外2条序列,其压缩效果提升最明显.

综上所述,尽管本文算法在编码时,会引入无效编码区编码代价,但借助希尔伯特空间填充,将基因组序列映射到二维进行编码有可能获得比直接进行一维序列建模更好的压缩效果.这也说明通过二维映射进行Context建模的方法对基因组序列压缩是可行的,因此达到我们最初设计目标.

(下转第65页)

式并非将课堂完全交给学生,教师就无所事事了.相反,为了达到较好的教学效果,对教师的要求不是降低了,反而是提高了.在第2阶段和第3阶段的教学过程中,教师需要提前预览各组的准备情况,提前与各组组长交流,确定课堂讲解的形式和内容;课堂的节奏也需要教师来把握,比如规定讨论多长时间、上台讲解多长时间等.另外,需要抽出额外的时间指导学生进行编程环境搭建、编程语言学习等.

3)以学生主导式为名,行教师讲授式之实.学生主导式的一个特点是学生学习的渐进性,即在防止冒险推进学生主导模式的同时,要注意学生讨论的盲目性、无序性.有的老师认识不到讨论课堂的效率提升是需要多次锻炼这一特点,在组织一两次讨论而达不到目的之后,又回到讲授式教学的起点,而只是象征性留几分钟让学生讨论.

只有任课教师对教学环境、教学内容、学生情况有了充分认识,并对可能出现的问题留有预案,才能真正推进学生主导式教学这一新模式,使课堂效率提升一个台阶.

4 小结

本文针对数学建模课程教学模式开展了讨论与探究.针对当前教学模式的优缺点和数学建模课程的教学特点,提出了三阶段学生主导模式的教学策略.文中除了详细讨论该模式下各阶段师生之间角色转换、授课与考核方式转变等问题外,还给出了该模式的教学安排实例以供授课教师参考.根据本人

多年的教学实践经验,该教学策略能够很好地实施以学生为主导的教学模式,对新一轮的高校数学教学改革具有一定的理论与实践意义.

[参考文献]

- [1]姜启源,谢金星.一项成功的高等教育改革实践:数学建模教学与竞赛活动的探索与实践[J].中国高教研究,2011(12):79-83.
- [2]黄廷祝,高建.大学数学研究型教学方法和考试方法改革与实践[J].中国大学教学,2012(11):52-55.
- [3]李大潜.关于高校数学教学改革的一些宏观思考[J].中国大学教学,2010(1):7-9.
- [4]李大潜.将数学建模思想融入数学类主干课程[J].中国大学教学,2006(1):9-11.
- [5]付军,朱宏,王宪昌.在数学建模教学中培养学生创新能力的实践与思考[J].数学教育学报,2007,16(4):93-95.
- [6]刘卫锋,何霞,王尚志.高中数学建模中教师问题初探[J].数学通报,2007,46(10):13-16.
- [7]樊士德,张维.高校经济类课程教学方法与效果评价比较研究[J].高等财经教育研究,2013,16(3):18-24.
- [8]孟祥林.互动课堂的困境与师生行为边界分析[J].宁波大学学报:教育科学版,2010,32(1):37-43.
- [9]孟祥林.影响因素与对策:基于博弈理论的高效教学过程分析[J].湖南师范大学教育科学学报,2007,6(2):47-51.
- [10]沈兴华,杨健荣,王成洲,等.教师应引导学生成为学习中的主导[J].南京军医学院学报,2000,22(1):43-44.
- [11]罗李平,杨柳,魏继东,等.关于数学建模教学与竞赛的思考[J].湖南工业大学学报,2010,24(1):94-96.

(上接第46页)

5 结论

本文中,我们给出一种基于希尔伯特空间填充的Context建模方法.通过将基因组序列映射到二维图像,并对映射图像编码,从而充分利用了碱基间的相关性.尽管通过二维映射,无效编码区代价被包含进最终压缩结果,但通过实验对比发现,这样的代价并不足以导致编码效率降低,相反,借助二维映射建模,基因组序列压缩效果相比前人方法略有提高.

[参考文献]

- [1] MATSUMOTO T, SADAKANE K, IMAI H. Biological sequence compression algorithms[J]. Genome Informatics, 2000, 11: 43-52.
- [2] DEOROWICZ S, GRABOWSKI S. Robust relative compression of genomes with random access[J]. Bioinformatics, 2011, 27(21): 2979-2986.
- [3] DEOROWICZ S, DANEK A. Genome compression: a novel approach for large collections[J]. Bioinformatics, 2013, 29(20): 2572-2578.
- [4] PINHO A J, NEVERS A, BASTOS C, et al. DNA coding using finite-

- context models and arithmetic coding, proceeding of ICASSP [J]. Acoustics, Speech and Signal Processing, 2009, 32: 1693-1696.
- [5] CAO M D, DIX T I, ALLISON L, et al. A simple statistical algorithm for biological sequence compression [C]//17th Data Compression Conference, 2007: 43-52.
- [6] PINHO A J, PRATAS D, FERREIRA P J. Bacteria DNA sequence compression using a mixture of finite-context models [J]. IEEE Statistical Signal Processing Workshop, 2011, 46: 125-128.
- [7] 陈旻,王开云,贾学明,等.基于加权Context建模的DNA序列压缩算法[J].昆明学院学报,2014,36(3):81-84.
- [8] 王笋,徐小双. Hilbert 曲线扫描矩阵的生成算法[J].中国图形图像学报,2000,11(1):119-122.
- [9] 陈旻,王开云,薛洁,等.一种图像自适应小波压缩算法[J].昆明学院学报,2013,35(6):96-99.
- [10] WU Xiao-lin, ZHAI Guang-tao, YANG Xiao-kang, et al. Adaptive sequential prediction of multidimensional signals with applications to lossless image coding[J]. IEEE Transactions on Image Processing, 2011, 20: 36-42.
- [11] NCBI. National center for biotechnology information [EB/OL]. [2014-10-05]. <http://www.ncbi.nlm.nih.gov/>.