doi:10.3969/j.issn.1672-5565.2014.02.09

一种新型的基于图像的DNA序列可视化模型

封海清,陆祖宏*
(东南大学生物科学与医学工程学院,南京 210018)

摘　要:传统的DNA序列可视化模型局限于短DNA序列的可视化,并且缺乏对可视化图形的通用分析方法。因此,文章提出了一种基于图像的DNA序列可视化模型,这种模型通过将一维的DNA序列转换为二维的256色的灰度图像,可以实现长DNA序列的可视化,具有很高的空间紧密性。借助成熟的图像处理方法来分析DNA可视化图像,可以获取原始DNA序列的规模、4种不同碱基的分布、无序程度等重要信息。通过比较不同DNA序列的可视化图像,可以获取这些序列的相似性信息。

关键词:DNA;可视化;图像处理
中图分类号:Q78　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-02-133-07

A novel visual modal of DNA sequence based on image

FENG Haiqing, LU Zuhong*
(*Department of Biomedical Engineering*, *Southeast University*, *Nanjing* 210018, *China*)

Abstract: Traditional visual modals of DNA sequence are limited to short DNA sequences and lack a general analyzing method of the visual graph. We put forward a novel visual modal of DNA sequence that transforms one dimensional DNA sequence into two dimensional 256-color gray-scale image, making the visualization of long DNA sequence possible. We can get the scale, distribution of four different bases, disorder of the original DNA sequence by analyzing the visual image of DNA sequence with the sophisticated image processing methods. We can also get the similarity between different DNA sequences by comparing their visual images.

Keywords: DNA; Visualization; Image process

人类基因组计划产生了大量的DNA数据,对这些DNA数据的可视化分析、压缩是非常重要的课题。传统上,DNA数据是以字符串形式保存的,这种格式缺乏直观性,同时占用比较大的内存空间。为了解决这两个问题,研究者提出了DNA可视化以及DNA压缩技术。

当前的DNA可视化模型大多是通过将4种碱基编码成空间的4个向量,通过DNA行走技术实现对DNA序列的可视化,比如非常著名的Gates-Nandy模型[1-2]。在Gates-Nandy模型的基础上,研究者为了解决退化以及信息丢失的问题,提出了CGR模型[3]、图细胞模型[4]、双碱基模型[5]、三维可视化模型[6-7]、蠕虫模型[8]等,并且基于一些可视化模型实现了不同DNA序列的比对[9-12]。当前的这些可视化模型的共同缺陷在于不适合对长DNA序列的编码。此外,在得到DNA序列可视化图形后,仅仅依靠肉眼对图形进行简单的分析,没有通用的方法对可视化图形进行分析和处理。

当前的DNA压缩技术分为无损和有损压缩,其中无损压缩技术尤为重要。随着近年来二代测序技术的快速发展,DNA的无损压缩技术得到了广泛的关注和研究[13-20]。当前的DNA无损压缩技术主要关注的是如何提高DNA序列的压缩率,来减少DNA数据存储需要的空间。研究者很难从压缩后的DNA数据中获取比较直观的DNA信息。

本文提出的DNA序列可视化模型通过对4种碱基进行编码,将DNA序列转换为256色的灰度图像。一方面具有极大的空间紧密型,实现了DNA数据在空间的压缩;另一方面,灰度图像中蕴含着非常丰富的信息,研究者可以借用非常成熟的图像处理

收稿日期:2014-01-06;**修回日期**:2014-02-22.
作者简介:封海清,男,硕士研究生,研究方向:生物信息学;E-mail:haiqingfeng@seu.edu.cn.
*通信作者:陆祖宏,男,教授,博士生导师,研究方向:基因测序方法,生物信息分析;E-mail:zhlu@seu.edu.cn.

方法来分析这些可视化图像,从中获取原始 DNA 序列中非常有用的信息。

1 算法原理

1.1 DNA 序列可视化模型评价指标

DNA 序列可视化模型的评价,可以通过下面 5 个指标来进行。

(1)退化:退化指的是在 DNA 序列可视化的结果图中出现圈,使得研究者无法确定 DNA 序列是顺时针还是逆时针的现象,从而研究者无法从结果图中还原出原来的 DNA 序列信息;

(2)信息丢失:信息丢失指的是 DNA 序列的可视化结果图与 DNA 序列间的关系不是一一对应的现象。这往往导致几条相似 DNA 序列的可视化结果图是相同的;

(3)可视化 DNA 序列的长度:目前已经测得的很多基因的长度达到了数万 bp,这就要求算法能够很好地适应长片段 DNA 序列的可视化;

(4)可视化图占用空间的大小:研究者为了便于观察和分析大量长片段 DNA 序列,要求可视化图占用的空间越小越好;

(5)可视化图能否反映有用信息:DNA 序列的可视化图应该可以明显地展示 DNA 序列信息的特征,尽可能多地反映序列中的有用信息。

1.2 基于灰度图像的 DNA 序列可视化模型

已有 DNA 可视化模型的一个共同缺陷在于不能对长片段的 DNA 序列进行可视化研究,特别是大于 10 万 bp 的 DNA 序列。本文提出了一种基于 256 色灰度图像的 DNA 序列可视化方案,可以较好地解决这个问题。

DNA 序列由 4 种核苷 A、T、C、G 排列组合而成。从信息论的角度来看,对 4 种碱基编码只需要两个 bit 就可以了。常见的编码方案是 A-00,T-11,C-01,G-10。

将 DNA 序列信息转化为图像信息后,可以利用非常成熟的图像处理技术对 DNA 序列信息进行分析。目前比较常用的图像是 256 色图,即 8 位图,考虑到上面提及的 4 种碱基的编码方案,本文将每 4 个碱基组合成一个 8 位的数据,这个数据的范围是 0~255。然后,将得到的一维数据转化为二维的图像信息。之后,将图像保存为 png 压缩格式的图像文件。

在具体实施的时候,有两个注意事项:(1)、需要在原始的 DNA 序列后面添加 3 个碱基“AAA”,保证每个 DNA 序列中的碱基都可以编码;(2)、假设原始 DNA 序列的长度为 L,那么寻找一个尽可能小的 N,使得 N * N>=L,在 DNA 序列编码成 1 维数据的后面补上(N * N-L)个 0,然后将 1 维数据转化成二维的图像。这样,任何一个 DNA 序列都可以编码得到宽高相等的可视化图像。

下面举例说明本文中提出模型的实施过程:假设一个 DNA 序列为“ATCGAACT”,首先在 DNA 序列后面补上 3 个碱基“AAA”,得到“ATCGAACTAAA”“ATCG”编码得到第一个像素点,“TCGA”编码得到第二个像素点,如此类推,“TAAA”编码得到第 8 个像素点,这 8 个像素点的灰度值分别为“54”、“216”、“96”、“129”、“7”、“28”、“112”、“192”,原始的 DNA 序列一共 8 位,L=8,所以满足 N * N>=8 的最小的 N 是 3,(N * N-L)= 1,因此在得到的一维数据的后面补上 1 个 0,得到新的一维数据,然后将这个一维数据转化为二维的图像,图像的大小为 3 * 3。

不难发现,保存为这种格式的图像文件后,研究者可以在很小的空间内可视化长片段的 DNA 序列,一幅大小为 1 024 * 1 024 的图像的可以存储的碱基的数量超过了 100 万。因此,这种编码方案是非常适合编码长片段的基因组序列的。

1.3 本文提出模型的优点

从评价 DNA 序列可视化模型的 5 个指标来看,本文中提出模型的优点:

(1)该模型是将一维数据一行一行地扫描为二维的图像数据的,因此,不存在圈的问题,也不会出现退化的现象;

(2)该模型中 DNA 序列与之得到的可视化图是一一对应的。在 DNA 序列的可视化过程中,没有任何信息的丢失;

(3)本文将一维的 DNA 序列信息转为二维的图像信息,大幅度压缩了 DNA 序列可视化的长度。采用传统的 DNA 序列可视化模型,几乎不可能对 100 万 bp的 DNA 序列进行可视化,而采用本文提出的模型,100 万 bp 的 DNA 的可视化,只需要 1 000 * 1 000的 256 色灰度图就可以实现了。实现长片段 DNA 序列的可视化也是这个算法的一大优点;

(4)从可视化图占据空间的角度来看,由于将 DNA 信息编码成 8 位的像素值,然后将像素值填满整个图像,一幅占用空间非常小的图像可以包含极为庞大的数据量,经过计算,1 幅大小只有 144 kb 的 png 图像可以容纳 50 万 bp 长的 DNA 序列信息;

(5)从能否反映有用信息的角度来看,在本文提出的模型中,我们可以从 DNA 序列的可视化图中获取如下信息:DNA 的序列规模、DNA 序列的 4 种

碱基的分布情况、DNA 序列的无序程度等,并且可以进行不同 DNA 序列间的相似性比较。

2　模型应用

正如上文所谈到的,DNA 序列的可视化结果图能否具有实际应用价值是非常重要的一个指标。本文将着重介绍从 DNA 序列可视化图中可以得到的信息。

2.1　DNA 序列可视化图像

从网站 http://www.ncbi.nlm.nih.gov/Ftp/下载了文件名为 hs_ref_GRCh37.p13_chrY.fa 的人类基因组的 Y 染色体数据。从文件中选取了一段长度为 160 000 bp 的 DNA 序列,序列从文件的第 10 288 509 bp到 10 448 508 bp,用 Seq1 表示该序列。图 1 展示了该 DNA 序列的可视化图像,图像的大小是 400 * 400。从这个图像中,我们可以看到图像中存在着一些重复结构。此外,我们可以从这个图像中看出,DNA 序列的规模是 400。

为了进行比较,下面采用两种经典的 DNA 序列可视化模型对这段 DNA 序列进行可视化。

DNA 序列光谱型二维可视化模型通过将 DNA 序列转化为二维的曲线,实现了 DNA 序列的可视化。可视化过程如下:首先画出 4 条平行线,代表 A、T、C、G,前面提到的序列 Seq1 的前面几个碱基分别是 GAAT…,因此,在 4 条平行线上分别标出这几个碱基对应的点。以此类推,所有碱基对应的点标好后,用线把它们连起来就可以了。图 2 展示了这个模型对 Seq1 的前 100 个碱基序列的可视化光谱曲线。图 3 展示了这个模型对 Seq1 的前 1 000 个碱基序列的可视化图谱曲线。

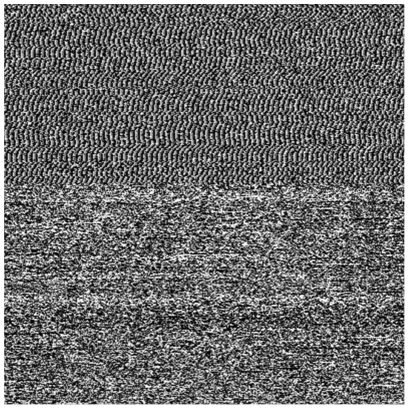


图 1　Seq1 的可视化图像

Fig.1　Visual image of seq1

从图 2 和图 3 中可以看到这个模型对于较短的 DNA 序列,能够反映一些 DNA 序列的性质,但是随着 DNA 序列长度的增加,可视化图谱上的曲线都挤到了一起,很难从视觉上分辨。因此,这个模型并不适合长 DNA 序列的可视化。

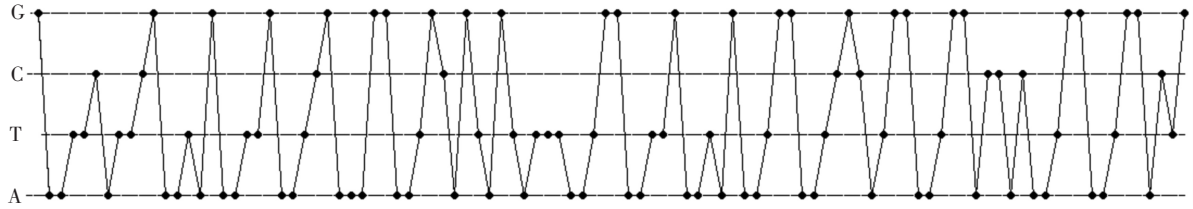


图 2　Seq1 前 100 个碱基的光谱曲线

Fig.2　Spectral curve of the first 100 bp of Seq1

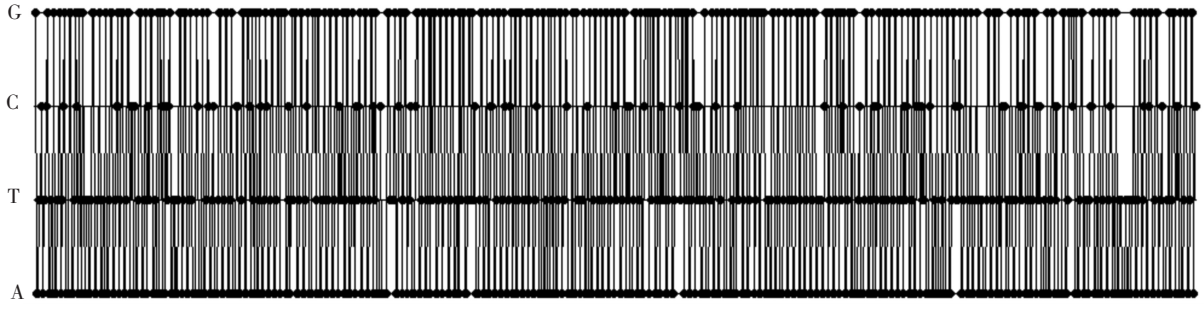


图 3　Seq1 前 1 000 bp 碱基的光谱曲线

Fig.3　Spectral curve of the first 1 000 bp of Seq1

DNA 序列双向量二维可视化模型也是一种比较常用的 DNA 序列可视化模型。这种模型采用了 DNA 行走技术,将 4 种碱基编码成两个方向的移动向量。编码方式如下:

A⇒(1,1)(1,1)
T⇒(1,1)(1,-1)
C⇒(1,-1)(1,1)
G⇒(1,-1)(1,-1)

把曲线的初始点设在原点,开始构造 DNA 序列的可视化曲线,对于 Seq1,它的前面几个碱基分别是 GAAT…,序列的第一个碱基是'G','G'编码的向量是(1,-1),(1,-1),所以,先后向(1,-1)(1,-1)两个方向分别移动一步,到达(2,-2),序列的第二个碱基是'A','A'编码的向量是(1,1),(1,1),所以先后向(1,1,),(1,1)两个方向分别移动一步,到达(4,0)。以此类推,完成整条 DNA 序列的"行走"。

图 4 展示了模型对于 Seq1 的前 10 000 bp 碱基序列的二维可视化曲线。从图中,我们可以看出随着 DNA 序列长度的增加,曲线的很多细节都遗失了,整个曲线变得模糊不清,只能看到一种趋势。

其他的传统的 DNA 序列可视化模型与这两种模型有相似的缺陷,即并不适合长 DNA 序列的可视化,空间紧密性低。并且可视化的曲线更依靠人眼的观察,没有一种比较通用成熟的分析工具。

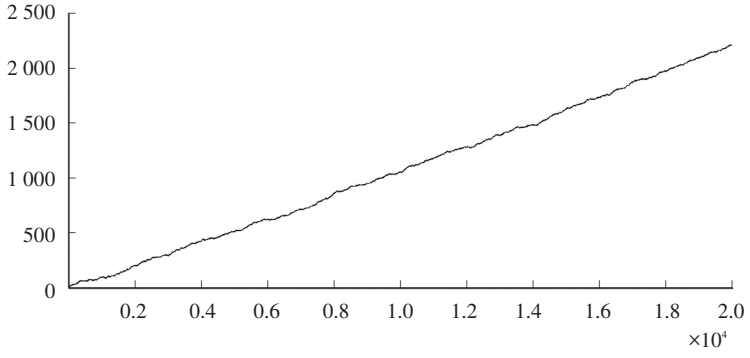


图 4 Seq1 的前 10 000 bp 碱基的二维向量可视化曲线

Fig. 4 2-D vector curve of the first 10 000 bp of seq1

2.2 DNA 序列可视化图像的直方图分析

通过直方图分析,我们可以获悉图像的总体的灰度分布情况。图 5 是图 1 中的可视化图像的直方图。根据前面提到的编码方案,在直方图中,[0,63]代表 A 的分布;[64,127]代表 C 的分布;[128,191]代表 G 的分布,[192,255]代表 T 的分布。考虑到在 DNA 编码图像的过程中,在一维数据的末端插入一些 0 来得到宽高相等的图像,这相当于在原始 DNA 序列的末端又插入了一些"A"。因此,在直方图中 A 碱基的含量比实际情况稍多一些。不过,我们依然可以从直方图中知道原始 DNA 序列中 4 种碱基的大致分布情况。

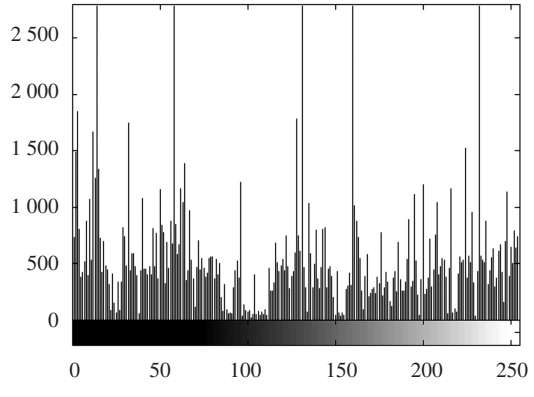


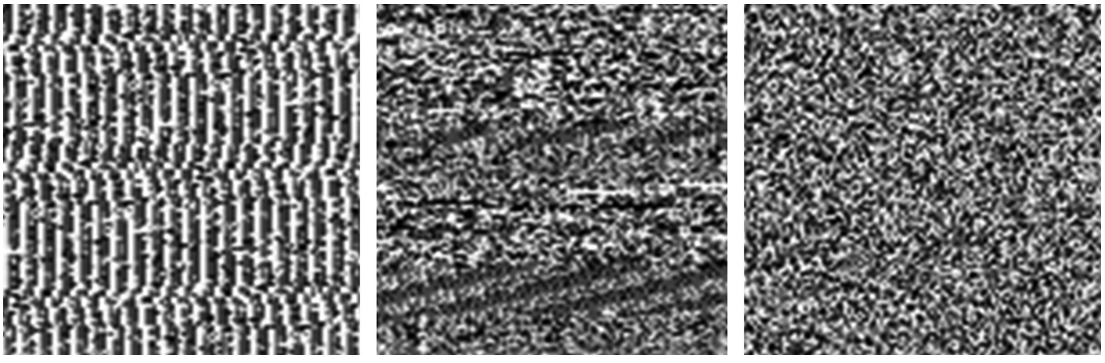
图 5 图 1 的直方图

Fig. 5 Histogram of Figure 1

2.3 DNA 序列可视化图像的熵分析

图像的熵反映了图像的无序程度。通常,图像的熵越大,表明图像的无序程度越高;相反,图像的熵越小,表明图像的无序程度越低。为了对比不同 DNA 序列的熵的情况,本文选取了 3 条不同的 DNA 序列,对它们的可视化图像进行了熵分析。序列来源于 2.1 节中提到的 Y 染色体数据文件,其中:序列 1是 Y 染色体数据文件中的一段 10 000 bp 长的 DNA 序列,序列从文件的第 25 259 292 bp 到 25 269 291 bp;序列 2 为 Y 染色体数据文件中的一段 10 000 bp长的 DNA 序列,序列从文件的第34 822 bp 到 44 821 bp;序列 3 为随机生成的一段10 000 bp长的 DNA 序列。这 3 种不同的 DNA 序列的可视化图见图 6,图 6(a)、(b)、(c)分别代表上面提到的 3 种不同的 DNA 序列。我们可以发现在图 6(a)中存在着一些重复的单元,图 6(c)则完全无序。

从直观的角度来看,从左到右,图像的无序度越来越高。为了更加精确地描述这 3 幅图像的无序度,本文计算了它们的熵,采用的图像熵的计算公式为:$E=-\sum_{i=0}^{255}P(i)*\log_2 P(i)$,其中 $P(i)$ 表示图像中灰度值为 i 的像素所占的比例。计算得到的结果见表 1。从表中,我们可以看出包含很多重复结构的图像的熵最小,而随机生成的碱基序列的熵最大,这与直观印象相符。



(a)　　(b)　　(c)

图 6　三种不同 DNA 序列的可视化图

Fig. 6　Visual images of three different DNA sequences

注:(a) 序列 1;(b) 序列 2;(c) 序列 3。

Notes:(a) Sequence1;(b) Sequence2;(c) Sequence3。

表 1　三种不同 DNA 序列的可视化图的熵

Table 1　Entropy of visual image of three different DNA sequences

	序列 1 编码图像	序列 2 编码图像	序列 3 编码图像
大小	100 * 100	100 * 100	100 * 100
熵	5.851 0	7.567 1	7.978 0

从网站 http://www.ncbi.nlm.nih.gov/Ftp/下载了包含人类的一部分基因的数据文件,文件名为 refseqgene1.genomic.fna。从中随机选取了1 000个基因,计算了这些基因序列可视化图像的熵,得到图 7 所示的结果,横轴表示可视化图像的宽度,纵轴表示可视化图像的熵。从图中可以看出,大部分基因序列可视化图像的熵在 7.7~7.8。

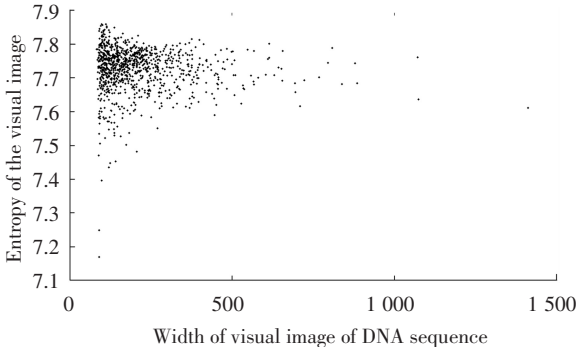


图 7　1 000 个基因的序列可视化图像的熵

Fig. 7　Entropy of visual image of 1 000 gene sequences

为了进行类比,本文分析了随机生成的 DNA 序列的可视化图像的熵(见图 8)。图中横坐标表示图像的宽度(高度与宽度相等),纵坐标表示图像的熵的大小。从图中可以明显看出,当图像的大小超过 50 * 50 的时候,图像的熵已经达到了 7.8,而当图像的大小达到 200 * 200 的时候,图像的熵已经非常接近于 8 了,序列越长,图像的熵越趋近于 8。

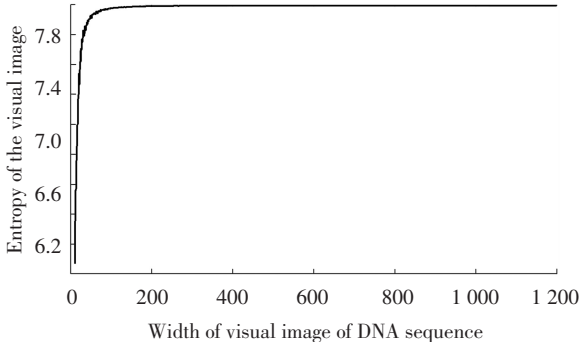


图 8　随机生成 DNA 序列的可视化图像的熵

Fig.8　Entropy of visual image of the random DNA sequences

对比图 7 和图 8,发现即使基因序列很长,基因序列编码图像的熵始终小于 7.9,而大部分的基因序列编码图像的熵局限在 7.7~7.8,这表明基因序列并非完全随机的,而是在无序中存在着某种规律性。

2.4　DNA 序列间相似度分析

DNA 序列间的相似性分析在生物信息学中是一个非常重要的课题,当前的很多相似性分析都是基于文本字符串实现的。正如前面所提到的,DNA 序列间的相似性分析也是 DNA 序列可视化图的一个重要应用场合。

如果两条 DNA 序列是相似的,那么它们的可视化图像也是相似的,DNA 序列间的相似性比较可以转换为两幅图像的相似性比较。比较两幅图像的相似度的方法有如下几种:基于两幅图像的欧氏距离、基于图像的直方图距离、基于 SVD (Sigular value decomposition)距离比较、基于图像特征点的比较等。下面简要介绍 SVD 在 DNA 序列相似性比较中的应用。

图像就是一个二维矩阵,奇异值分解(SVD)是基于整体的表示,具有稳定性和抗噪性,能够非常好地代表图像的特征,这种方法通常比简单的基于图

像灰度值判断图像间的相似度的可靠性要高。

从网站 http://www.ncbi.nlm.nih.gov/Ftp/下载了包含很多人类 mRNA 的数据文件,文件名为 human.rna.fna。从中选取了 PNMA5 的 4 条转录变异序列,从 Homo sapiens paraneoplastic Ma antigen family member 5 (PNMA5), transcript variant 1 到 Homo sapiens paraneoplastic Ma antigen family member 5 (PNMA5), transcript variant 4,共 4 段序列,下面分析这 4 段序列间的相似性。

这 4 段序列的长度分别为 3 118、3 218、3 204、3 203,根据本文的算法模型,这 4 段序列编码得到的图像的大小分别是 56 * 56, 57 * 57, 57 * 57, 57 * 57。为了方便进行比较,在序列 1 的后面补上“AAA…”,使得编码得到的 4 个图像的大小都是 57 * 57。设图像矩阵为 $\boldsymbol{X}$,矩阵大小为 $m * n$。对图像矩阵进行 SVD 分解,基本算法如下:

(1)、求出 $\boldsymbol{X}^{\mathrm{H}}\boldsymbol{X}$ 的全部非零特征值 λ_i,则 $\sigma_i = \sqrt{\lambda_i}$ 为 $\boldsymbol{X}$ 的正奇值,$i = 1, 2, \cdots, r$,记 $\Delta = \mathrm{diag}(\sigma_1, \sigma_2, \ldots \sigma_r)$;

(2)、求酉矩阵 $\boldsymbol{U} \in \boldsymbol{U}^{m\times m}$,使得 $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{U} = \mathrm{diag}(\sigma_1{}^2, \sigma_2{}^2 \cdots, \sigma_r{}^2, 0, \ldots, 0)$;

(3)、设 $\boldsymbol{U} = [U_1, U_2]$, $V_1 = \boldsymbol{X}^{\mathrm{H}}\boldsymbol{U}_1\Delta^{-\mathrm{H}}$,则 V_1 为次酉阵,求 $V_2 \in \boldsymbol{U}_{n-r}^{n\times(n-r)}$,使得 $\boldsymbol{V} = [V_1, V_2] \in \boldsymbol{U}^{n\times n}$;

(4)、得到 $\boldsymbol{X}$ 的 SVD 分解:$\boldsymbol{X} = \boldsymbol{U}\begin{bmatrix}\Delta & O\\ O & O\end{bmatrix}\boldsymbol{V}^{\mathrm{H}}$。

Matlab 集成了上述的算法,成为一个专门的函数 SVD。采用 matlab 对上述的图像矩阵进行 SVD 分解,matlab 代码如下:[$\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$] = svd($X$)。其中,$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 是酉矩阵,大小分别为 $m * m, n * n$,$\boldsymbol{S}$ 是对角矩阵,大小为 $m * n$。矩阵 $\boldsymbol{X}$、$\boldsymbol{U}$、$\boldsymbol{S}$、$\boldsymbol{V}$ 满足:$\boldsymbol{X} = \boldsymbol{U} * \boldsymbol{S} * \boldsymbol{V}'$。设两个图像矩阵分别是 $X1$, $X2$,对 $X1$ 和 $X2$ 进行 SVD 分解,得到两个反映图像特征的对角矩阵 $S1$、$S2$。取 $S1$ 和 $S2$ 的对角线元素,得到 $D1$ 和 $D2$ 两个一维向量。设 $\mathrm{d}D = D1 - D2$,则 $\mathrm{d}D$ 的 2 范数为 $\| \mathrm{d}D \|_2 = \sqrt{\sum_{i=0}^{i=n-1} (\mathrm{d}D(i))^2}$,其中 $\mathrm{d}D(i)$ 表示向量 $\mathrm{d}D$ 的第 i 个分量。如果这个 2 范数越大,表明两个图像相似度越小,如果这个 2 范数越小,表明两个图像的相似度越高。表 2 显示了计算的结果,表中的数值表示两个特征向量差的 2 范数。

从表 2 中,我们可以清楚地看到,transcript variant 3 与 transcript variant 4 的相似性最好,transcript variant 1 与其它三条序列的相似性最差。采用本文提出的可视化模型,可以从图像特征的角度来分析序列的相似性,给研究者提供了一个新的工具和视角。

表 2 四条 mRNA 序列的相似性比较

Table 2 Similarity comparison of 4 mRNA sequences

	Transcript variant 1	Transcript variant 2	transcript variant 3	transcript variant 4
transcript variant 1	0	207.756 2	218.828 4	199.878 8
transcript variant 2	207.756 2	0	119.646 7	140.898 6
transcript variant 3	218.828 4	119.646 7	0	112.050 8
transcript variant 4	199.878 8	140.898 6	112.050 8	0

3 总结和展望

本文提出了一种新型的基于图像的 DNA 序列可视化模型,可以实现长 DNA 序列的可视化,具有很高的空间紧密性,实现了 DNA 数据在空间的压缩。通过观察和分析 DNA 序列的可视化图像,研究者可以更加直观地研究和分析 DNA 序列。借助于成熟的图像处理手段,研究者可以从图像中获取原始 DNA 序列的规模、DNA 序列中不同碱基的分布、DNA 序列的无序度,并且可以进行不同 DNA 序列间的比较。

本文提出的可视化模型将传统的 DNA 可视化与 DNA 压缩技术有机地结合了起来,得到可视化图像后,研究者可以根据自己的需要,自主选择成熟的图像处理方法,对 DNA 可视化图像进行处理和压缩。

参考文献(References)

[1] GATES M A. A simple way to look at DNA[J]. Journal of Theoretical Biology, 1986, 119(3): 319-328.

[2] NANDY A. A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to glob in genes [J]. Current Science, 1994, 66(4): 309-314.

[3] MILAN R. Another look at the chaos-game representation of DNA[J]. Chemical Physics Letters, 2008, 456: 84-88.

[4] YAO Yuhua Yao, WANG Tianming. A class of new 2-D graphical representation of DNA sequences and their application[J]. Chemical Physics Letters, 2004, 398(4-

6): 318-323.

[5] DAI Qi, LIU Xiaoqing, WANG Tianming. A novel 2D graphical representation of DNA sequences and its application [J]. Journal of Molecular Graphics and Modelling, 2006, 25(3): 340-344.

[6] ARAM V, IRANMANESH A. 3D-Dynamic representation of DNA sequences [J]. MATCH Communications in Mathematical and in Computer Chemistry, 2012, 67: 809-816.

[7] NAFISEH J, IRANMANESH A. C-curve: A novel 3D graphical representation of DNA sequence based on codons[J]. Mathematical Biosciences, 2013, 241(2): 217-224.

[8] MILAN R. Graphical representations of DNA as 2-D map [J]. Chemical Physics Letters, 2004, 386: 468-471.

[9] HUANG Guohua, LIAO Bo, LI Yongfan, et al. H-Lcurve: A novel 2D graphical representation for DNA sequences[J]. Chemical Physics Letters, 2008, 462(1-3): 129-132.

[10] ZHANG ZhuJin. DV-Curve: A novel intuitive tool for visualizing and analyzing DNA sequences [J]. Bioinformatics, 2009, 25(9): 1112-1117.

[11] HUANG Guohua, LIAO Bo, LI Yongfan,et al. Similarity studies of DNA sequences based on a new 2D graphical representation[J]. Biophysical Chemistry, 2009, 143(1-2): 55-59.

[12] LIAO Bo, XIANG Qilin, CAI Lijun, et al. A new graphical coding of DNA sequence and its similarity calculation[J]. Physica A, 2013, 392(19): 4663-4667.

[13] KENNY D, PAUL R, SCOTT C, et al. Data structures and compression algorithms for high-throughput sequencing technologies[J]. BMC Bioinformatics, 2010, 11: 514-526.

[14] CONGMAO W, DABING Z. A novel compression tool for efficient storage of genome resequencing data[J]. Nucleic Acids Research, 2011, 39(7): e45.

[15] MUHAMMAD N S, TANG Jijun, ZHENG W J, et al. Improving transmission efficiency of large sequence alignment/Map (SAM) files[J]. PLoS One, 2011, 6(12): e28251.

[16] ARMANDO J P, DIOGO P, SARA P G.GReEn: A tool for efficient compression of genome resequencing data[J]. Nucleic Acids Research, 2012, 40(4): e27.

[17] MARKUS H F, RASKO L, GUY C, et al. Efficient storage of high throughput DNA sequencing data using reference-based compression [J]. Genome Research, 2011, 21: 734-740.

[18] DANIEL C J, WALTER L R, PENG Xinxia, et al. Compression of next-generation sequencing reads aided by highly efficient de novo assembly[J]. Nucleic Acids Research, 2012, 40(22): e171.

[19] MARK H. High-Throughput compression of FASTQ data with SeqDB [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2013, 10(1): 213-218.

[20] POPIISCH N,HAESELEL A V. NGC: lossless and lossy compression of aligned high-throughput sequencing data [J]. Nucleic Acids Research, 2013, 41(1): e27.