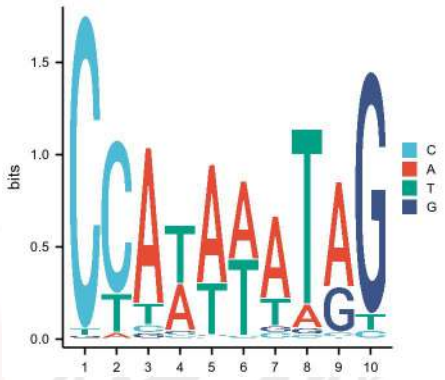
基础绘图 - 序列分析图-字符串



网址：https://www.xiantao.love

更新时间：2023.06.26

目录

基本概念

- 序列分析图：sequence logo，序列通常指的是核苷酸(在 DNA/RNA 链中)或氨基酸(在蛋白质序列中)。每个位置出现的碱基或氨基酸类型反映了该位置序列的偏好性，每个字母的大小与该碱基在该位置上的出现频率成正相关。

应用场景

在生物信息分析中，常使用序列分析图(sequence logo)来直观清晰的反应序列偏好特征，如突出序列比对中的保守位置，用于研究结构域序列相似性；临床上可视化 DNA、RNA 和蛋白质结合位点(激酶，SH2 / SH3 域，转录因子（TFs），RNA 结合蛋白，核酸酶，核糖核蛋白等)探索突变对重大疾病的影响。

分析过程

- 数据格式：（具体数据格式要求可以看后面过程的“数据格式”部分）
  - 数据提供每一列含有等长的序列信息的表格，至少 1 列，如下：

MA0002.1
AATTGTGGTTA
ATCTGTGGTTA
AATTGTGGTAA
TTCTGCGGTTA
AATTGCGGTAA
TATTGCGGTTT
TATTGTGGTAG
TTTTGTGGTAG
ATATGTGGTAA
AACTGCGGTTG
GACTGTGGTTG
CTTTGCGGTTA
ATCTATGGTTA
CATTGCGGTTT
TATTGTGGCTA
TCTTGTGGTAT
TGCTGCGGTTA

- 数据清洗：对所选择的某一列进行数据清洗，每一行必须是由等长的字符串构成，如 motif 序列；自动删除含有空值的行。
- 数据处理：
  - 每一行字符串切割为单个字符（如碱基），每一个字符代表序列中的一个位置，构成字符矩阵。

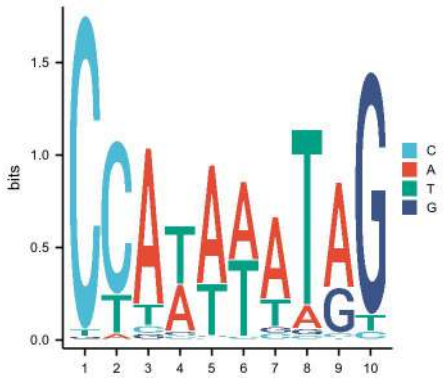
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
[1,]	A	A	T	T	G	T	G	G	T	T	A
[2,]	A	T	C	T	G	T	G	G	T	T	A
[3,]	A	A	T	T	G	T	G	G	T	A	A
[4,]	T	T	C	T	G	C	G	G	T	T	A
[5,]	A	A	T	T	G	C	G	G	T	A	A
[6,]	T	A	T	T	G	C	G	G	T	T	T
[7,]	T	A	T	T	G	T	G	G	T	A	G
[8,]	T	T	T	T	G	T	G	G	T	A	G
[9,]	A	T	A	T	G	T	G	G	T	A	A
[10,]	A	A	C	T	G	C	G	G	T	T	G

- 由字符矩阵统计每个位置中不同字符的个数，最后通过不同的方式计算字符高度（见下方计算方式）。

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
A	10	12	4	1	2	2	0	0	0	8	13
T	11	11	14	24	1	16	0	0	25	16	7
G	3	1	1	0	23	0	26	26	0	0	4
C	2	2	7	1	0	8	0	0	1	2	2

- 计算方式：
  - **经典计数**（custom）：直接使用频数数据作为各字符串的高度绘图。
  - **概率**（probability）：对统计个数数据第 2 列以后的数值计算各字符串的占比（每一列计算频率）绘图。
  - **香农熵**（bits）：在计算频率矩阵的基础上计算每一列的香农熵，香农熵作为系数与概率矩阵相乘后，获得的数据作为各字符串的高度进行绘图。

主要结果



- 上图，纵坐标为根据香农熵计算（具体计算见 分析过程）字符高度，横坐标为对应的位置（单个字符对应的位置）
- 不同颜色代表字符类型（第 1 列）
- 字符默认按照从小到大向上堆叠

数据格式

	A	B	C	D	E	F
1	MA0001.1	MA0002.1	MA0004.1	MA0005.1	MA0006.1	MA0007.1
2	CCATATATAG	AATTGTGGTTA	CACGTG	CCTAATTGGGC	CGCGTG	AAAAGTACACCCTGTACCGACA
3	CCATATATAG	ATCTGTGGTTA	CACGTG	CCTAATTTGGC	CGCGTG	CTAAGCACACCGTGTCCCAGTC
4	CCATAAATAG	AATTGTGGTAA	CACGTG	CCTAATCGGGC	CGCGTG	TTAAGAACACTCTGTACGACAC
5	CCATAAATAG	TTCTGCGGTTA	CACGTG	CCTAATCGGGC	CGCGTG	AGTAGAACATAATGTTCCGACA
6	CCATAAATAG	AATTGCGGTAA	CACGTG	CCTAATTGGGA	CGCGTG	GTAAGTACACTCTGTTCCGACA
7	CCATAAATAG	TATTGCGGTTT	CACGTG	CCAAATGGGGT	CGCGTG	TAAGGAACACCCTGACCCCCCT
8	CCATAAATAG	TATTGTGGTAG	CACGTG	CCAAATGGGGT	CGCGTG	CCTAGAACGTCCTGATCCGCCC
9	CCATATATGG	TTTTGTGGTAG	CACGTG	CCTAATCGGGT	CGCGTG	AAGAGAACATCACGTTCCTATG
10	CCATATATGG	ATATGTGGTAA	CACGTG	CCTAATTAGGC	TGCGTG	TAAAGCACGACGCGTTCTCAAC
11	CCAAATATAG	AACTGCGGTTG	CACGTG	CCTAATTAGGC	TGCGTG	ATAGGCACATGGTGTACAATCG
12	CCAAATATAG	GACTGTGGTTG	CACGTG	CCAAATAGGGT	TGCGTG	AAAATTACATCGTGAACCCTCG
13	CCAAATATAG	CTTTGCGGTTA	CACGTG	CCTAATTCGGT	TGCGTG	ATGAGAACTTAGAGTACCCAAC
14	CCATAAATGG	ATCTATGGTTA	CACGTG	CCTAAATTGGC	TGCGTG	CGGAGCACAGCCTGTCCCTGCA
15	CCATAAATGG	CATTGCGGTTT	CACGTG	CCAAATTAGGT	TGCGTG	ACTAGCACACAGTGTACTACAT
16	CCATAAATGG	TATTGTGGCTA	CACGTG	CCTAATCTGGA	TGCGTG	CGTAGAACGAAATGTTCTCTCC
17	CCATAAATGG	TCTTGTGGTAT	AACGTG	CCTAATTCGGA	TGCGTG	CTCAGCACAGTCCGTACCTGAC
18	CCAAAAATAG	TGCTGCGGTTA	AACGTG	CCAAATATGGT	TGCGTG	AACGGACCGTCCTGTACACGAC
19	CCAAAAATAG	TTTTGTGGTCT	AACGTG	CCAAATATGGT	TGCGTG	CAAGGATCACTATGAACCTGAC
20	CCAAAAATAG	GCTTGTGGTTT	AACGTG	CCAAATATGGA	TGCGTG	GTTAGTACGCGGCGTTCCGGAG
21	CCAAAAATAG	AAATGTGGTAC	CGCGTG	CCCAATTGGGA	AGCGTG	TTTGGGACACAGGGTACCTCCG
22	CCAAAAATAG	TTATGAGGTTA		CCATATTTGGC	AGCGTG	TACATAACGTCCAGTCCCTGCG
23	CCAAATATGG	GTCTGCGGTAT		CCATATTGGGT	GGCGTG	AAAGGAACATCATGTGAAATAA
24	CCAAAAATGG	AAAAGTGGTTA		CCTAATACGGT	GTCGTG	TTACGTACGATACGTAGCCGAC
25	CCAAAAATGG	TTGTATGGTTT		CCTAATTCGGG	AGTTTG	CTTGGACCGATGAGTCCTAGTC
26	CCATTTATAG	AATTGAGGTCC		CCTAAAATGGC		
27	CCATTTATAG	TTTCTTGGTTA		CCTAATTAGGG		
28	CCATTTATAG			CCGAATGGGGT		
29	CCATTAATAG			CCAAAACGGGA		
30	CCATTAATAG			CCTAAACCGGC		

数据要求：

- 数据至少 1 列，1 行。每一行必须为等长的字符串(可以是由碱基或氨基酸字母、0~9 之间的任意数字够长)
- 只允许包含字母[大小写 a~z]或者数字[0~9]范围内的字符。
- 最多支持 30 列，5000 行。若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。

参数说明

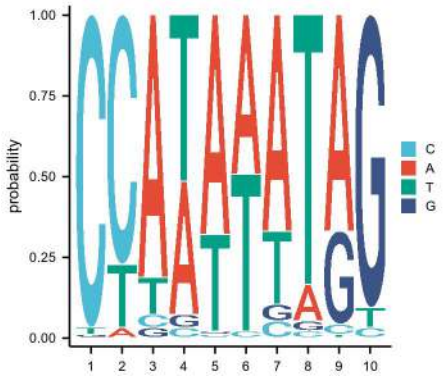
（说明：标注了颜色的为常用参数。）

变量列表

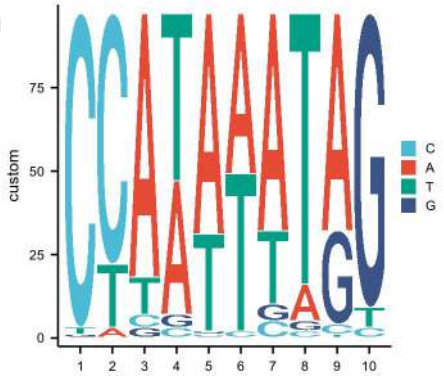
- 绘图变量：本模块可通过选取文件表格的列名，选择需要绘图的数据范围。该参数的选项值由上传文件的列名进行动态调整，注意，必须有列名信息。

方法

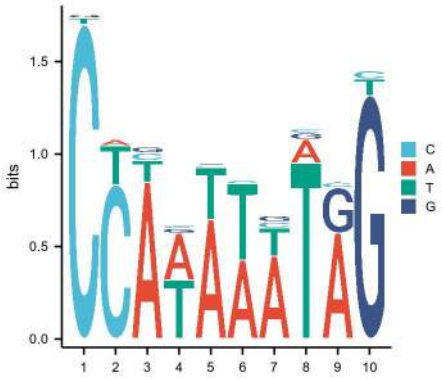
- 高度计算：字符高度的计算方式，默认为香农熵，可选择香农熵、概率、经典计数，具体计算方式见 分析过程。
  - 计算-概率，纵坐标表示各字符类型在各位置中的占比



  - 计算-经典计数，纵坐标表示各字符类型在各位置中的计数



- 字符上下顺序颠倒：主要影响字符的排列方式，默认按照高度值从小到大向上堆叠。
  - 计算-香农熵，字符上下顺序颠倒



堆栈

- 颜色：字符的填充色颜色选项，有多少个字符会提取多少个颜色，最多支持修改 10 个颜色。受配色方案全局性修改。
- 不透明度：字符的宽度控制。默认 0.95，范围设置在 0~1 之间。
- 不透明度：字符的透明度。0 为完全透明，1 为完全不透明。

标题

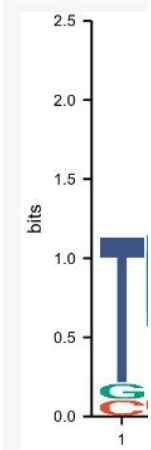
- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]。

图注(Legend)

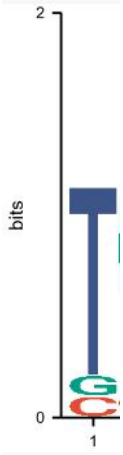
- 是否展示：是否展示图注
- 图注标题：可以添加图注标题
- 图注位置：可选择 默认、右、上、下。

坐标轴

- 是否显示 x 轴：默认展示 x 轴
- 是否显示 y 轴：默认展示 y 轴
- x 轴标注旋转：支持对 x 轴文字进行旋转。适合于 x 轴文字过长的时候
- y 轴范围+刻度：（注意：范围的修改如果调整过大会失效）
  - 如果只是想要修改范围，可以只输入两个范围值，比如 0,2.5，自动调整刻度



- 如果同时想要修改范围+刻度，可以输入比如：0,0,2,2 。注意，此时最大和最小值会被当做范围值，不会作为刻度，如果需要刻度，需要类似于 0.5 那样同时写两次



风格

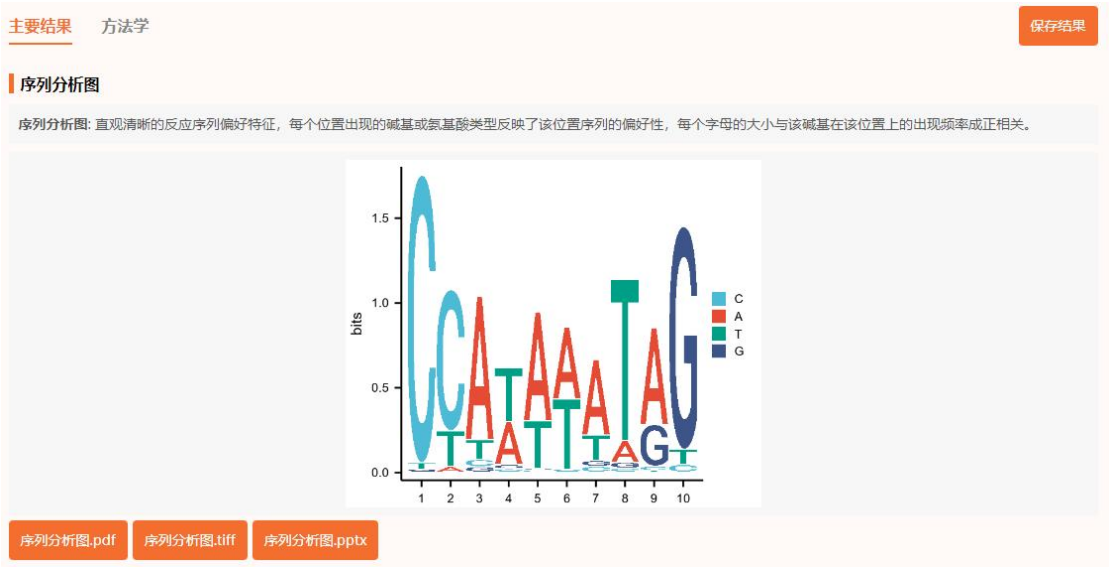
- 边框：可以选择是否进行添加图形边框的操作
- 网格：可以选择是否进行添加图形网格线的操作
- 文字大小：控制整体文字大小，默认为 6pt

图片

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果



主要结果格式为图片格式，提供 PDF、TIFF 和 PPTX 格式下载。

方法学

方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包：ggseqlogo（用于计算）、ggplot2 包（用于可视化）

处理过程：将清洗后的数据用 ggseqlogo 包处理，再用 ggplot2 包绘制序列分析图。

如何引用

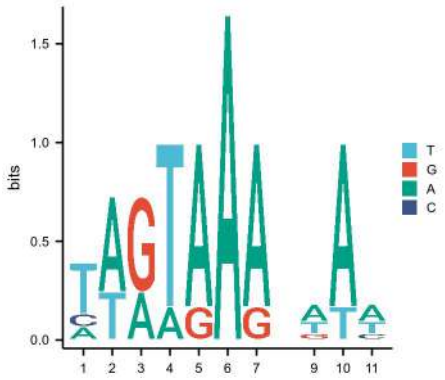
生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. 为什么图中位置空了？明明有字符。

答：



这是由于该位置计算香农熵时出现负数导致的，ggseqlogo R 包中计算香农熵（bits）的过程中，如果存在负数将会被替换为 0，这是该 R 包内部的处理方式，无法调整；此时，可以选择其他计算方式来绘图。