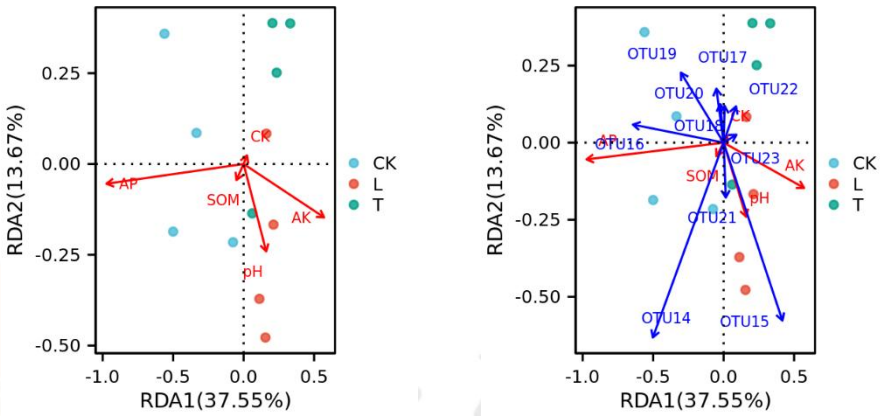
表达差异 – 聚类-RDA/CCA 分析



网址：https://www.xiantao.love

更新时间：2023.11.08

目录

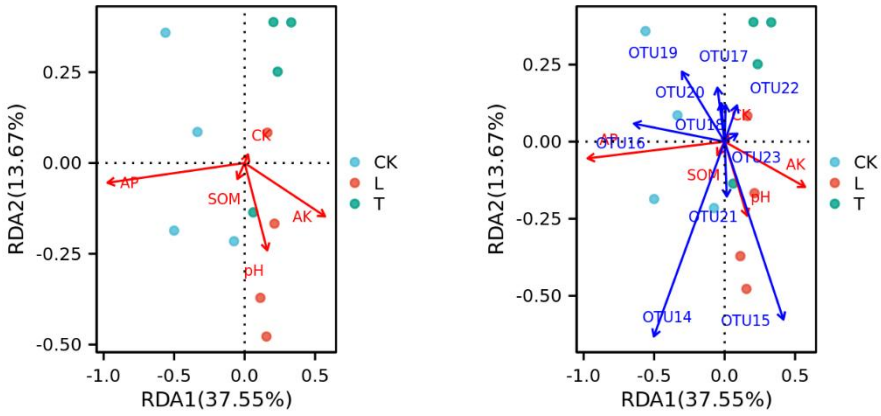
基本概念

- RDA/CCA 分析：冗余分析(redundancy analysis,RDA)或者典范对应分析(canonical correspondence analysis, CCA)，可以用来分析环境因子对群落分布影响。

应用场景

- 在研究微生物群落时，除了比较不同群落样品的差异外，还需要了解哪些因素会对微生物群落产生影响。
  - 可以将微生物群落数据与环境因子数据相结合，进行 RDA / CCA 分析。
  - 也可以用来分析代谢物、菌群、样品三者之间的关系，通过整合多种不同类型的数据，可以利用 RDA/CCA 分析方法揭示复杂生物系统中的关联和交互作用，为揭示生物系统的整体特征提供重要信息。

主要结果



对于主要结果的解释：

- RDA/CCA 的结果图中使用点代表不同的样本，不同颜色的点属于不同分组，两点之间的距离越接近，说明两个样品的菌群组成/功能相似度越高；
- 从原点发出的红色箭头代表不同的环境因子，箭头的长度代表该环境因子对群落变化影响的强度，箭头的长度越长，表示该环境因子的对菌群组成/功能的影响越大；
- 红色箭头与坐标轴的夹角代表该环境因子与坐标轴的相关性，夹角越小，代表相关性越高；
- 样本点到环境因子箭头及其延长线的投影距离表示环境因子对样本的影响强度，样本点与箭头距离越近，该环境因子对样本的作用越强；
- 样本位于箭头同方向，表示环境因子与样本物种群落的变化正相关，样本位于箭头的反方向，表示环境因子与样本物种群落的变化负相关；
- 图像中坐标轴标签中的数值，代表了坐标轴所代表的环境因子组合对物种群落变化的解释比例。

数据格式

物种数据

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	# group	CK	CK	CK	CK	L	L	L	L	T	T	T	T
2	OTU_id	CK1	CK2	CK3	CK4	L1	L2	L3	L4	T1	T2	T3	T4
3	OTU1	32	7	0	5	16	7	61	25	7	16	26	2
4	OTU2	2	1	0	0	27	14	46	15	27	11	14	24
5	OTU3	0	0	0	1	19	23	29	19	0	25	23	4
6	OTU4	0	1	0	0	2	0	1	0	19	33	19	72
7	OTU5	0	1	0	0	0	2	17	2	23	8	10	68
8	OTU6	8	29	33	17	0	0	2	1	0	0	0	0
9	OTU7	0	7	0	0	0	0	0	1	49	5	55	13
10	OTU8	1	0	0	0	4	1	0	0	38	18	20	35
11	OTU9	2	0	1	0	30	14	55	7	1	0	1	0
12	OTU10	30	6	7	6	11	0	0	0	3	11	31	16
13	OTU11	0	0	1	0	6	0	54	25	0	3	2	2
14	OTU12	1	0	0	0	2	1	0	1	30	11	6	43
15	OTU13	0	0	0	0	0	0	0	0	4	26	36	17
16	OTU14	321	411	505	424	356	345	375	350	99	181	242	170
17	OTU15	7	209	52	25	317	237	161	229	49	88	30	123
18	OTU16	216	149	147	200	25	12	24	2	2	12	7	2

数据要求：

- 头部注释行（以#开头）：
  - 用于表征每个样本所属的分组，至少需要提供 1 行样本的注释信息，每行的分组最多是 10 个。注意，注释行不能超过 4 行。
- 主体部分：
  - 数据至少有 4 列（除第 1 列 OTU id）以上，至少需要 3 行(除样本名行)数据。
  - 主体的第一行为样本编号（如图中的第 2 行），这一行不能含有缺失、重复及特殊字符。
  - 主体的第一列为变量名（示例是微生物多样性测序数据中的 OTU id，也可以是 ASV id 或物种比如界门纲目等）。
  - 主体的其他部分为样本在各个维度对应的数值，不能含有非数值内容。
  - 数据不能含有整列之和为 0 的样本。
  - 数据中不能含有负数值。
- 最多支持 100 列（除第 1 列 OTU id），5000 行(除样本名行)。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。

环境数据

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	# group	CK	CK	CK	CK	L	L	L	L	T	T	T	T
2	Samples	CK1	CK2	CK3	CK4	L1	L2	L3	L4	T1	T2	T3	T4
3	AK	8.44	8.45	8.46	8.47	8.58	8.49	8.7	8.51	8.52	8.53	8.54	8.55
4	pH	1.54	1.63	1.57	1.87	1.6	1.67	1.67	1.88	1.63	1.57	1.59	1.87
5	SOM	30.43	31.3	30	65.84	29.76	31.89	31.5	57.98	30.89	29.47	30.06	59
6	AP	77.27	47.99	64.3	90.46	34	35	36.82	38.06	39.44	30.69	29.2	26.65
7	CK	40.22	40.95	37.43	81.63	34.93	39.73	38.57	80.31	39.53	35.53	36.15	88.74

数据要求：

- 头部注释行（以#开头）：
  - 用于表征每个样本所属的分组，至少需要提供 1 行样本的注释信息，每行的分组最多是 10 个。注意，注释行不能超过 4 行。
- 主体部分：
  - 数据至少有 4 列（除第 1 列环境因子）以上，至少需要 3 行(除样本名行)数据。
  - 主体的第一行为样本编号（如图中的第 2 行），这一行不能含有缺失、重复及特殊字符。
  - 主体的第一列为变量名（示例是环境因子，也可以是代谢物等）。
  - 主体的其他部分为样本在各个维度对应的数值，不能含有非数值内容。
  - 数据不能含有整列之和为 0 的样本。
  - 数据中不能含有负数值。
- 最多支持 100 列（除第 1 列环境因子），100 行(除样本名行)。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。
- 环境因子数目必须小于样本数。（当环境因子的数量大于等于样本数时，模型参数的估计就会变得不稳定，可能无法得出有效的参数估计值）。

参数说明

（说明：标注了颜色的为常用参数。）

物种数据参数

- 转换：物种数据处理参数，可选择 无、非 0 数值以 2 为底对数变换后+1、每个数值除以样本列总和、每个数值除以物种行最大值、先平方根转化再弦转化、有无数据（1-0）

  默认无就是物种数据不做处理，如选择其他处理参数则由 decostand 函数进行计算。

统计分析

- 算法：可选择 auto、RDA、CCA，默认 auto。

  auto 会根据 DCA 分析得到的 Axis lengths 的第一轴的大小选择合适的方法（小于 3.0，默认 RDA；在 3.0-4.0 之间，选 RDA 和 CCA 都可以，auto 默认 RDA 分析；大于 4.0，默认选择 CCA）。

- Mantel 相关性方法：补充结果中群落与环境因子相关性分析（Mantel Test）的相关性方法。

- 种子号：设置种子数可以保证统计检验分析结果可重复，默认为 2023，此参数请输入非零整数。

点

- 填充色：点的填充色颜色选项，取决于上传数据中的头部注释行信息，有多少个分组会提取多少个颜色，最多支持修改 10 个颜色。受配色方案全局性修改。
- 描边色：点的描边色颜色选项，取决于上传数据中的头部注释行信息，有多少个分组会提取多少个颜色，最多支持修改 10 个颜色。受配色方案全局性修改。
- 样式：点的样式类型，可选择 圆形、正方形、菱形、三角形、倒三角，默认为圆形。多选，多选后不同的分组/分类中的点的类型也会有相应变化，循环取该参数值。
- 大小：点的大小。
- 不透明度：点的透明度。0 为完全透明，1 为完全不透明。

样本外圈

- 展示：是否需要圈住分组的不同分类。
- 样式：外圈的样式类型，可选择连线、椭圆，默认为连线。单选，选择类型后所有圈的样式都统一改变。
  - 椭圆，即置信椭圆。（注意，不是所有的分类都能有圈的，如果分类内含有极端的样本，可能没有办法有圈，另外样本多少也会影响是否能有圈，如单个分组内少于 4 个样本则无法添加）。
  - 连线，是由各个组最外层的点连接而成，起码两个样本及以上。
- 线条类型：外圈的线条类型，可选择 实线、虚线，默认为虚线。单选，选择类型后所有圈的描边都统一改变。
- 线条粗细：外圈的线条粗细，默认为 0.75pt。

环境因子-标注

- 类型选择：是否需要展示环境因子信息。可选择 不展示、展示全部环境因子、展示下面特定环境因子，默认为展示全部环境因子。（注意只是展示，环境因子都参与分析）
- 特定环境因子：当上一个参数选择了“标注下面特定环境因子”时，将根据此参数输入的环境因子在图上进行标注，一行一个。注意环境因子是否与上传的环境数据保持一致！
- 线条颜色：环境因子箭头颜色选项。不受配色方案全局性修改。
- 线条类型：环境因子箭头的线条类型，可选择 实线、虚线，默认为实线。单选，选择类型后所有环境因子箭头的线条都统一改变。
- 线条粗细：环境因子箭头的线条粗细，默认为 0.75pt。
- 不透明度：环境因子箭头的透明度。0 为完全透明，1 为完全不透明。

物种-标注

- 类型选择：是否需要展示物种信息。可选择 不展示、展示 top10 丰度物种、展示下面特定物种，默认为不展示。（注意只是展示，物种都参与分析）
- 特定物种：当上一个参数选择了“标注下面特定物种”时，将根据此参数输入的物种在图上进行标注，一行一个。注意物种是否与上传的物种数据保持一致！
- 线条颜色：物种箭头颜色选项。不受配色方案全局性修改。
- 线条类型：物种箭头的线条类型，可选择 实线、虚线，默认为实线。单选，选择类型后所有物种箭头的线条都统一改变。
- 线条粗细：物种箭头的线条粗细，默认为 0.75pt。
- 不透明度：物种箭头的透明度。0 为完全透明，1 为完全不透明。

样本-标注

- 类型选择：是否需要标注样本编号信息。可选择 不标注、标注全部样本、标注下面特定样本，默认为不标注。
- 特定样本：当上一个参数选择了“标注下面特定样本”时，将根据此参数输入的样本编号在图上进行标注，一行一个。注意样本编号是否同时存在于上传的物种数据和环境数据中（注意，只保留物种数据和环境数据中共有的样本进行分析）！
- 标注大小：控制图中需标注的文字大小，默认为 5pt。

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。选择经典类型风格时如果需要上标，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。
- 图注标签：可以修改图注中分组标签的名字，如果有多个名字要修改，则需要把这些名字以逗号的形式合并成一个，类似 A, B。
- 图注位置：可选右、上，至少有两个分组时默认为右。

风格

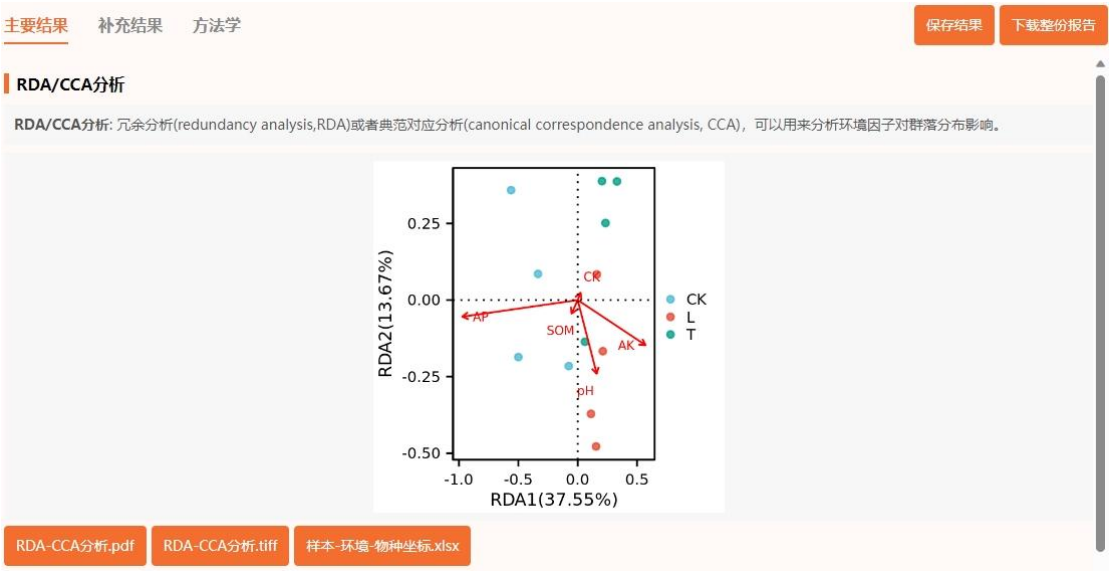
- 坐标样式：无边框的情况下，坐标轴的样式。可选择经典类型、指向类型，默认为经典类型。
- 边框：是否添加外框，默认添加。
- 网格：是否添加网格。
- 文字大小：针对图中所有文字整体的大小控制，默认为 7pt。

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 格式下载，结果报告可以下载包括 pdf 以及说明文本的内容。

	A	B	C	D
1	sample	RDA1	RDA2	
2	CK1	-0.5613848	0.35853633	
3	CK2	-0.07524216	-0.21547752	
4	CK3	-0.33445111	0.08543252	
5	CK4	-0.49992006	-0.18615405	
6	L1	0.16149758	0.08357054	
7	L2	0.15488864	-0.47778813	
8	L3	0.11186127	-0.37140863	
9	L4	0.21174495	-0.16699734	
10	T1	0.0611945	-0.1360341	
11	T2	0.20454075	0.38793639	
12	T3	0.23446748	0.25135357	
13	T4	0.33080297	0.38703042	

样本坐标 环境坐标 物种坐标

	A	B	C	D
1	env	RDA1	RDA2	
2	AK	0.57133304	-0.14704958	
3	pH	0.16029904	-0.24061074	
4	SOM	-0.05124216	-0.04431997	
5	AP	-0.97300098	-0.05433247	
6	CK	0.02393278	0.02379398	
7				
8				
9				
10				
11				
12				
13				

样本坐标 环境坐标 物种坐标

	A	B	C	D
1	species	RDA1	RDA2	
2	OTU1	0.01262833	-0.02884125	
3	OTU2	0.06917311	-0.04030784	
4	OTU3	0.06064185	-0.02590692	
5	OTU4	0.08784618	0.16905688	
6	OTU5	0.07317799	0.08360271	
7	OTU6	-0.06195329	-0.01846632	
8	OTU7	0.05024508	0.0496124	
9	OTU8	0.05411912	0.08206672	
10	OTU9	0.03241847	-0.09903169	
11	OTU10	-0.01409182	0.11037148	
12	OTU11	0.03296278	-0.08446414	
13	OTU12	0.0477041	0.07256958	

样本坐标 环境坐标 物种坐标

另外，提供样本-环境-物种坐标 xlsx 下载。

补充结果

去趋势对应分析（Detrended Correspondence Analysis，DCA）

在进行排序分析之前，我们需要先对物种群落数据进行去趋势对应分析 (DCA)，结果中根据Axis lengths的第一个值选择排序分析模型 。

DCA1	DCA2	DCA3	DCA4
2.2494	0.99342	0.9184	0.70333

去趋势对应分析结果.xlsx

通过DCA分析得到4个Axis lengths的数值，第一个值小于3.0，建议选RDA(基于线性模型，冗余分析)

此表格提供去趋势对应分析（Detrended Correspondence Analysis，DCA）结果 xlsx 下载。

通过 DCA 分析得到 4 个 Axis lengths 的数值。

第一个值小于 3.0，建议选 RDA(基于线性模型，冗余分析);

第一个值在 3.0-4.0 之间，选 RDA(基于线性模型，冗余分析)和 CCA(基于单峰模型，典范对应分析)都可以，如不指定分析方法 auto 会选择 RDA 分析;

第一个值大于 4.0，建议选 CCA(基于单峰模型，典范对应分析)。

解释因子解释率

解释因子对应的解释数据变异情况的比例以及累积比例情况

解释因子	解释比例(%)	累积比例(%)
RDA1	37.546	37.546
RDA2	13.672	51.218
RDA3	3.1984	54.416
RDA4	2.423	56.839
RDA5	1.1073	57.947

此表格为解释因子对应的解释数据变异情况的比例以及累积比例情况，如 RDA1 的解释比例为 37.546%。

单环境因子显著性检验（Envfit test）

环境因子	r2	pvalue
AK	0.3162	0.1683
pH	0.052607	0.8119
SOM	0.0026633	0.9901
AP	0.83707	0.0099
CK	0.00061955	1.0000

· r2 列代表相应环境因子的重要性度量值；
· pvalue 列代表相应环境因子的显著性水平。

此表格为单环境因子显著性检验（Envfit test）结果，揭示了每个环境因子与群落结构关联的显著性。

总体差异显著性（ANOVA like permutation test）

统计检验方法	统计量(pseudo_F)	显著性(P)
ANOVA like permutation test	1.6535	0.1089

· 统计量 pseudo_F值越接近于1，越不显著，pseudo-F值远大于1，则显著程度越高；
· 显著性 p 值，小于0.05表示统计具有显著性。

此表格为总体差异显著性（ANOVA like permutation test）结果，样本数过少可能导致没有这个分析结果。

群落与环境因子相关性分析（Mantel Test）

Mantel 分析是一种用于比较两个距离矩阵（或相似度矩阵）之间的相关性的统计方法。它的主要目的是检验两个距离矩阵是否有显著的相关性，从而评估它们之间的相似性或差异性。

统计检验方法	统计量(R)	显著性(P)
Mantel Test	0.11741	0.1782

· 统计量 R 为菌群与环境因子的相关性，R 值在 -1 和 1 之间
· 显著性 p 值，小于0.05表示统计具有显著性

此表格为群落与环境因子相关性分析（Mantel Test）结果，在这个分析中首先计算了物种数据和环境数据的距离矩阵，然后评估了环境因子的变化与微生物群落结构之间的相关性。

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：ggplot2 包（用于可视化）、vegan（用于分析）

处理过程：

⑴ 清洗后的物种数据和环境数据使用 rda.default 或 cca.default 函数进行 RDA/CCA 分析。

⑵ 使用 ggplot2 包对结果进行可视化。

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. 在进行 RDA/CCA 分析时为什么有时选择标注一些环境因子或物种箭头在图中显示不全?

答:

- 首先确定标注的环境因子或物种一定存在于上传的环境数据或物种数据中如果还是标注不全可能的原因包括:
  - 数据缺失:环境数据或物种数据中可能存在缺失值,导致某些环境因子或物种无法进行标注。
  - 共线性:当环境因子之间存在高度相关性时,RDA/CCA 分析可能会自动剔除其中一些环境因子,以避免多重共线性问题。这可能导致一些环境因子箭头无法显示。
  - 箭头重叠:在某些情况下,物种之间或物种与环境因子之间可能存在强烈的相关性,导致箭头重叠或无法完全显示。这可能会使部分箭头在图中显示不全。