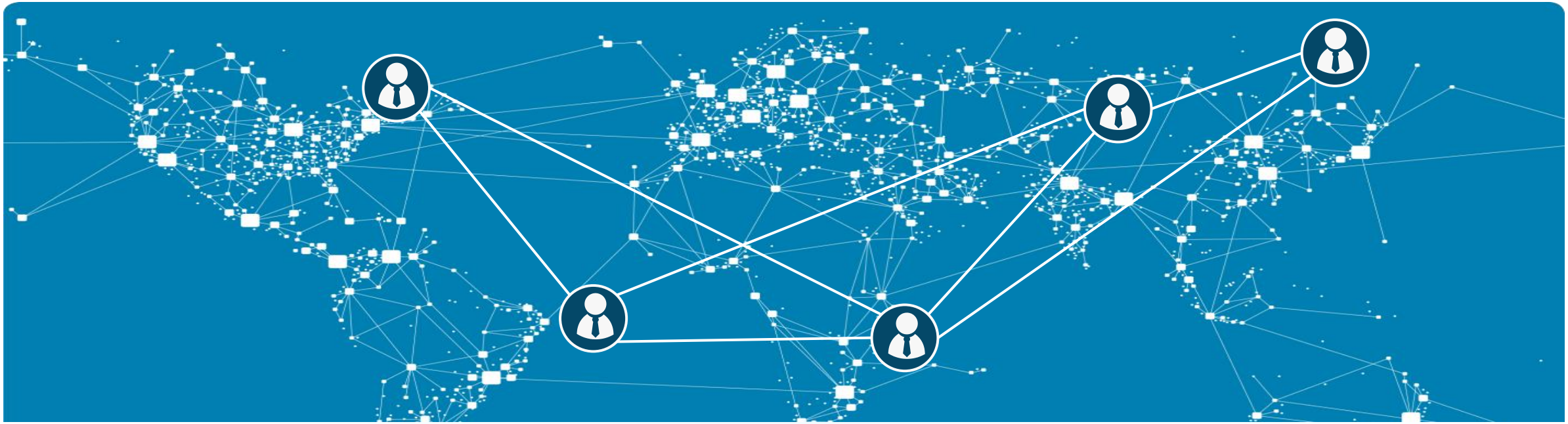
2020微生物组—扩增子16S分析研讨会第10期



36溯源和马尔可夫链实战

王金锋　　中科院北京生科院

2020年12月6日

目录

SourceTracker

- 工具网址：https://github.com/danknights/sourcetracker

Published in final edited form as:
Nat Methods. ; 8(9): 761–763. doi:10.1038/nmeth.1650.

Bayesian community-wide culture-independent microbial source tracking

Dan Knights[1], Justin Kuczynski[2], Emily S. Charlson[3,4], Jesse Zaneveld[2], Michael C. Mozer[1], Ronald G. Collman[3], Frederic D. Bushman[3], Rob Knight[5,6], and Scott T. Kelley[7]

SourceTracker数据输入格式

- 脚本实现：SourceTracker.Rmd
- 输入数据：

1.样本信息表

2.OTU丰度表

A	B	C	D	E	F
#SampleID	Description	Env	SourceSink	Study	Details
Run20100430_H2O-1	PCR water 1	Lab 1	sink	Lab 1	PCR_water_sample_1_2010_04_30_run
Run20100430_H2O-2	PCR water 2	Lab 1	sink	Lab 1	PCR_water_sample_2_2010_04_30_run
Run20100430_H2O-3	PCR water 3	Lab 1	sink	Lab 1	PCR_water_sample_3_2010_04_30_run
Run20100430_H2O-4	PCR water 4	Lab 1	sink	Lab 1	PCR_water_sample_4_2010_04_30_run
BF1	Alfisol 1	Soil	source	88_Soils	NA
CA2	Alfisol 2	Soil	source	88_Soils	NA
CO2	Alfisol 3	Soil	source	88_Soils	
DF3	Alfisol 4	Soil	source	88_Soils	
HI4	Andisol 1	Soil	source	88_Soils	
HJ2	Andisol 2	Soil	source	88_Soils	
JT1	Aridisol 3	Soil	source	88_Soils	
MD5	Aridisol 4	Soil	source	88_Soils	NA

样本所属类别
Source：来源
Sink：溯源目标

样本ID号

样本描述信息

样本ID号

A	B	C	D	E
# QIIME-formatted OTU table				
#OTU ID	X2d1h	X3d2w	X3d7d	X3d1h
Streptococcus	13848.35359	22595.64891	22952.09406	3396.990231
Neisseria	12784.38959	10592.64816	11563.50209	3537.798117
Prevotella	4453.509399	3428.357089	17170.34302	5825.926252
Haemophilus	5845.493923	510.1275319	4635.1180	2216844
Rothia	2275.1167	720.180045	7	6.916307
Veillonella	1614.870264	1590.397599	9134.1949	6.660213
Fusobacterium	8982.715842	1860.465116	991.1573219	1672.093637
Actinomyces	9924.723496	3090.772693	3585.657371	11942.26877
Capnocytophaga	11426.04819	32618.15454	2089.204159	13068.73185
Porphyromonas	857.8998276	292.5731433	1496.453212	79.20443545
Leptotrichia	7801.000883	2198.049512	1904.576815	20214.73202
Granulicatella	899.9537407	2303.075769	1370.129239	167.2093637
Bifidobacterium	0	7.501875469	281.7996307	8.800492828

Sourcetracker开头
两行使用带#号表头

OTU编号或
细菌分类名

相对丰度

使用R语言读取数据

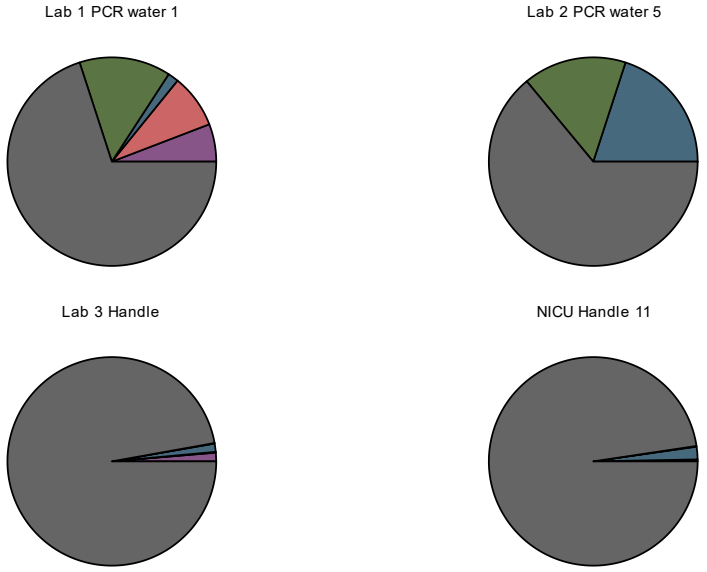
```
## 读取数据
>metadata = read.table('metadata_sourcetracker_trim.txt',sep='\t',h=T,row.names=1,check=F,comment= '' )
>otus = read.table('otu_sourcetracker.txt',sep='\t', header=T,row.names=1,check=F,skip=1,comment='')
> otus = t(as.matrix(otus))
> common.sample.ids = intersect(rownames(metadata), rownames(otus))      ## 提取otu表与metadata共有id
> otus = otus[common.sample.ids,]
> metadata = metadata[common.sample.ids,]       ## 从otu表中提取具有metadata信息的样本
> if(length(common.sample.ids) <= 1) {
   message = paste(sprintf('Error: there are %d sample ids in common '),
                   'between the metadata file and data table')
  stop(message)}
```

使用R语言进行溯源分析

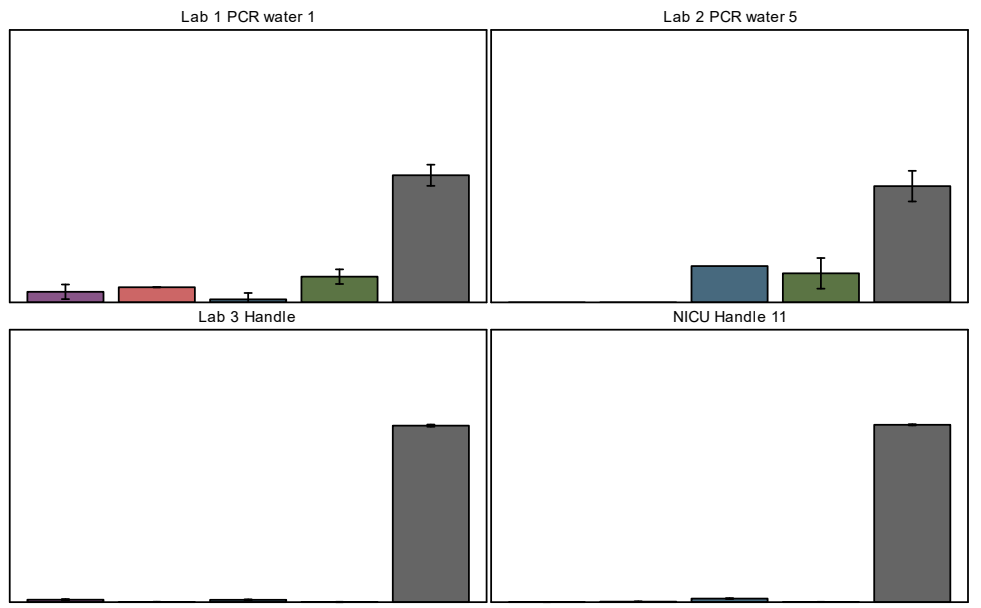
```
> train.ix = which(metadata$SourceSink=='source')    ## 提取训练集
> test.ix = which(metadata$SourceSink=='sink')    ## 提取测试集
> envs = metadata$Env    ## 提取环境信息
> if(is.element('Description',colnames(metadata))) desc = metadata$Description    ##提取样本标签
> source('SourceTracker.r')    ## 载入SourceTracker包
> tune.results = tune.st(otus[train.ix,], envs[train.ix])    ## 使用交叉验证计算alpha值
> alpha1 = tune.results$best.alpha1
> alpha2 = tune.results$best.alpha2
> st = sourcetracker(otus[train.ix,], envs[train.ix])    ## 使用训练集训练sourcetracker
> results = predict(st,otus[test.ix,], alpha1=alpha1, alpha2=alpha2)    ## 计算各来源比例
```

使用R语言进行可视化

- > labels = sprintf(‘%s %s’ , envs,desc) ##展示样本标签
- ## 展示SourceTracker溯源分析结果
- ## 可选择多种不同展示形式
- > plot(results, labels[test.ix], type='pie’)



> plot(results, labels[test.ix], type='bar')



目录

FEAST

- 工具网址：https://github.com/cozygene/FEAST

nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0431-x

FEAST: fast expectation-maximization for microbial source tracking

Liat Shenhav[1], Mike Thompson[2], Tyler A. Joseph[3], Leah Briscoe[2], Ori Furman[4], David Bogumil[4], Itzhak Mizrahi[4], Itsik Pe'er[3] and Eran Halperin[1,2,5,6*]

FEAST数据输入格式

- 脚本实现：FEAST.Rmd
- 输入数据：

1.样本信息表

2.OTU丰度表

A	B	C	D	E	F
#SampleID	Description	Env	SourceSink	Study	Details
Run20100430_H2O-1	PCR water 1	Lab 1	sink	Lab 1	PCR_water_sample_1_2010_04_30_run
Run20100430_H2O-2	PCR water 2	Lab 1	sink	Lab 1	PCR_water_sample_2_2010_04_30_run
Run20100430_H2O-3	PCR water 3	Lab 1	sink	Lab 1	PCR_water_sample_3_2010_04_30_run
Run20100430_H2O-4	PCR water 4	Lab 1	sink	Lab 1	PCR_water_sample_4_2010_04_30_run
BF1	Alfisol 1	Soil	source	88_Soils	NA
CA2	Alfisol 2	Soil	source	88_Soils	NA
CO2	Alfisol 3	Soil	source	88_Soils	
DF3	Alfisol 4	Soil	source	88_Soils	
HI4	Andisol 1	Soil	source	88_Soils	
HJ2	Andisol 2	Soil	source	88_Soils	
JT1	Aridisol 3	Soil	source	88_Soils	
MD5	Aridisol 4	Soil	source	88_Soils	NA

样本所属类别
Source：来源
Sink：溯源目标

样本ID号

样本描述信息

样本ID号

A	B	C	D	E
# QIIME-formatted OTU table				
#OTU ID	X2d1h	X3d2w	X3d7d	X3d1h
Streptococcus	13848.35359	22595.64891	22952.09406	3396.990231
Neisseria	12784.38959	10592.64816	11563.50209	3537.798117
Prevotella	4453.509399	3428.357089	17170.34302	5825.926252
Haemophilus	5845.493923	510.1275319	4635.1180	2216844
Rothia	2275.1167	720.180045	7	6.916307
Veillonella	1614.870264	1590.397599	9134.1949	6.660213
Fusobacterium	8982.715842	1860.465116	991.1573219	1672.093637
Actinomyces	9924.723496	3090.772693	3585.657371	11942.26877
Capnocytophaga	11426.04819	32618.15454	2089.204159	13068.73185
Porphyromonas	857.8998276	292.5731433	1496.453212	79.20443545
Leptotrichia	7801.000883	2198.049512	1904.576815	20214.73202
Granulicatella	899.9537407	2303.075769	1370.129239	167.2093637
Bifidobacterium	0	7.501875469	281.7996307	8.800492828

相对丰度

OTU编号或细菌分类名

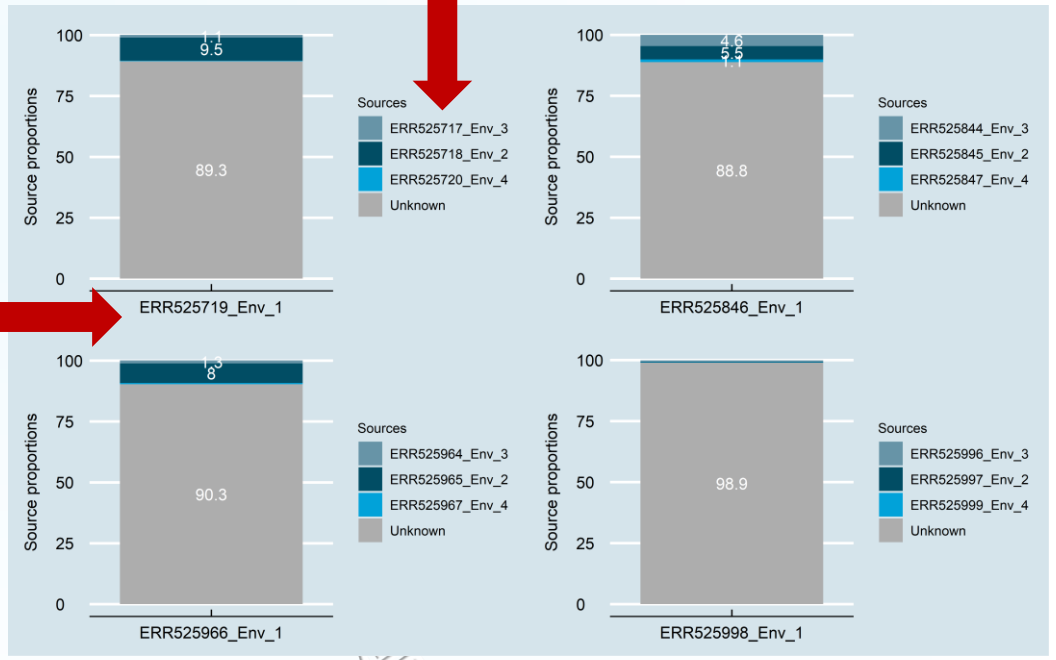
使用R语言进行溯源和可视化

易生信

```
## 读取数据
> metadata = Load_metadata(metadata_path = "metadata_example_FEAST.txt")
> otus = Load_CountMatrix(CountMatrix_path = "otu_example_FEAST.txt")
> dir_path = getwd() ## 设置结果输出路径
## 使用FEAST进行溯源
> FEAST_output = FEAST(C = otus, metadata = metadata,
different_sources_flag = 1,  dir_path = dir_path,outfile="demo")
## 展示FEAST溯源分析结果
> PlotSourceContribution(SinkNames = rownames(FEAST_output),
        SourceNames = colnames(FEAST_output), dir_path = dir_path,
        mixing_proportions = FEAST_output,
        Plot_title = "Test_",Same_sources_flag = 0, N = 4)
```

潜在来源

溯源目标



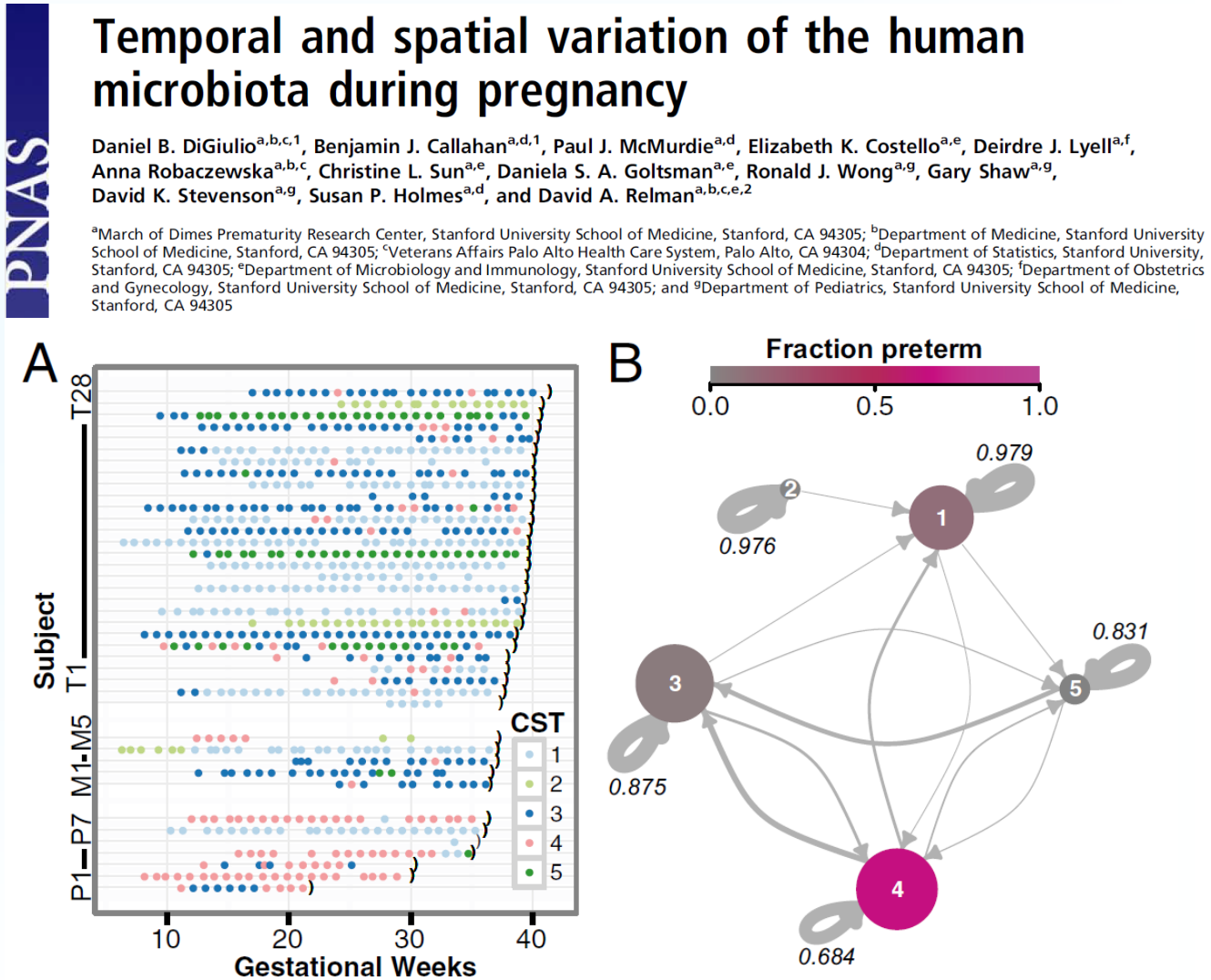
FEAST展示不同环境样本中潜在来源所占比例

目录

- SourceTracker
- FEAST
- 马尔可夫链

马尔可夫链测试数据

- 以DiGiulio et al. 2015年发表在*PNAS*上的阴道菌群数据为例
- 使用R语言对状态转移矩阵进行计算，并对马尔可夫链进行可视化



PNAS

Temporal and spatial variation of the human microbiota during pregnancy

Daniel B. DiGiulio[a,b,c,1], Benjamin J. Callahan[a,d,1], Paul J. McMurdie[a,d], Elizabeth K. Costello[a,e], Deirdre J. Lyell[a,f], Anna Robaczewska[a,b,c], Christine L. Sun[a,e], Daniela S. A. Goltsman[a,e], Ronald J. Wong[a,g], Gary Shaw[a,g], David K. Stevenson[a,g], Susan P. Holmes[a,d], and David A. Relman[a,b,c,e,2]

[a]March of Dimes Prematurity Research Center, Stanford University School of Medicine, Stanford, CA 94305; [b]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305; [c]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304; [d]Department of Statistics, Stanford University, Stanford, CA 94305; [e]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305; [f]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA 94305; and [g]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305

马尔可夫链数据输入格式

- 脚本实现：MarkovChain.Rmd
- 输入数据：包含不同时间点信息的信息列表
- 必需列：样本编号、个体编号、时间点、状态/分组信息

	样本编号	个体编号	时间点		状态/分组
Rowname	SampleID	SubjectID	Time	Group	CST
1000301298	1000301298	10003	198	7	A
1000301308	1000301308	10003	205	7	A
1000301318	1000301318	10003	212	8	A
1000301328	1000301328	10003	219	8	A
1000301338	1000301338	10003	226	8	A
1000401368	1000401368	10004	264	9	C
1000401378	1000401378	10004	271	10	C
1000501278	1000501278	10005	188	7	C
1000501308	1000501308	10005	211	8	C
1000501318	1000501318	10005	218	8	C
1000501368	1000501368	10005	258	9	C
1000501378	1000501378	10005	265	9	C

使用R语言计算概率转移矩阵

- >data_markov = read.table("data_markov.txt",header = T,sep = '\t',row.names = 1) ##数据读取
- >data_markov$SubjectID = as.character(data_markov$SubjectID)
- >data_markov$CST = as.factor(data_markov$CST)
- >CSTs = levels(data_markov$CST)
- >nstates = nlevels(data_markov$CST)
- >data_markov_prev = samdat.prune_prev(data_markov) ##调用转换周期函数
- >rownames(data_markov_prev) = data_markov_prev$SampleID
- >data_markov_prev$PrevCST = data.frame(data_markov)[data_markov_prev$PrevID,"CST"] ##提取每个样本前一个时间点的状态
- >data_markov_prev$CurCST = data_markov_prev$CST ##提取每个样本当前时间点状态
- >ttab = table(data_markov_prev$PrevCST, data_markov_prev$CurCST) ##计算不同状态间的转换频数
- >trans = matrix(ttab, nrow=nstates) ##计算概率转移矩阵
- >trans = trans/rowSums(trans) ##标准化至1

使用R语言进行可视化

```
>plotMC(mcPreg,
     edge.arrow.size = 0.8, edge.arrow.width = 1,  ##设置箭头大小与宽度
     edge.label = edge.labels, edge.label.font = 2, edge.label.cex = 0.8,   ##设置边标签字体与大小
     edge.label.color = "black",          ##设置边标签颜色
     edge.width = (15*wts + 0.1)*0.6, edge.curved = sapply(elcat, function(x) x %in% elrev),   ##设置边宽度与弯曲度
     edge.color = "#a3a3a3",       ##设置边颜色
     layout = layout, edge.loop.angle = edge.loop.angle,
     vertex.size = vert.sz,   vertex.color =  “#a9303b” ,     ##设置点大小与颜色
     vertex.label.font = 2,  vertex.label.cex = c(2,.5,2,2,1),      ##设置点标签字体与大小
     vertex.label.color = vert.font.clrs, vertex.frame.color = NA)
```

R语言输出文件

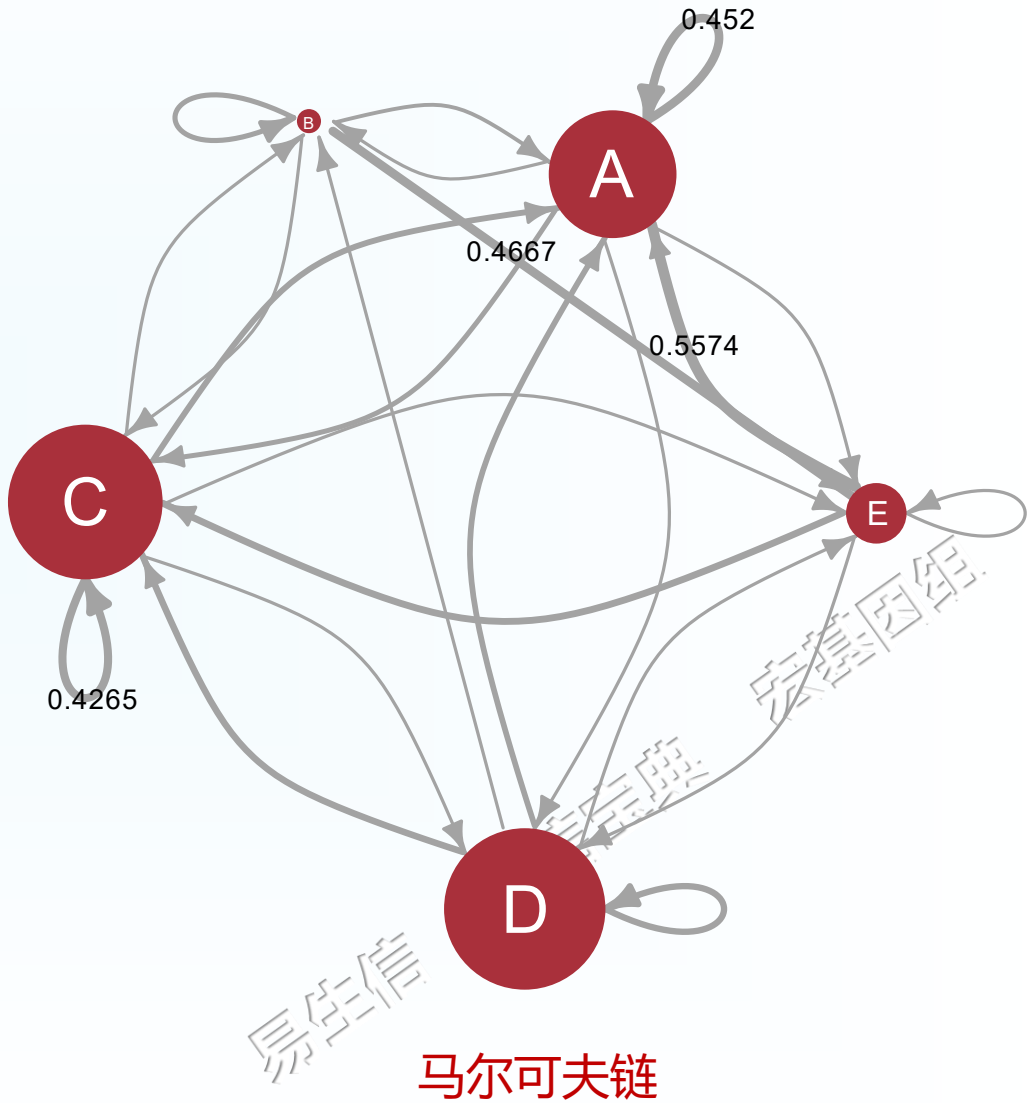
- 输出结果：

1.状态转移矩阵

2.马尔可夫链

```
> CSTtrans
          A          B          C          D          E
A 0.4520000 0.05200000 0.26000000 0.09600000 0.14000000
B 0.2000000 0.26666667 0.06666667 0.00000000 0.46666667
C 0.3088235 0.07352941 0.42647059 0.15686275 0.03431373
D 0.3000000 0.01111111 0.34444444 0.33333333 0.01111111
E 0.5573770 0.00000000 0.36065574 0.03278689 0.04918033
```

状态转移矩阵



马尔可夫链

总结

- 1.使用SourceTracker或FEAST进行溯源分析时，对尽可能多的潜在环境进行溯源会得到更好的效果
- 2.涉及人体跨位点或跨代传递的溯源，配对数据是最佳选择
- 2.溯源分析涉及多重复的时间序列时，可以根据需要对同一个时间点的溯源结果取用平均值或中位数
- 3.使用马尔可夫链进行状态转移分析时，状态的划分不宜过多，以3-5个为最佳

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识