R を用いたクラスター分析

清水顕史

データの準備

代謝産物×サンプルの行列データをタブ区切りのテキストファイルで用意します（データは Excel で準備し、タブ区切りで保存します)。例えば代謝産物×8 サンプルのメタボローム データを解析します。

	A	B	C	D	E	F	G	H	I
1		cont_12d	A_12d	B_12d	C_12d	cont_15d	A_15d	B_15d	C_15d
2	m1	0.0125	0.014	0.016	0.018	0.0135	0.015	0.014	0.018
3	m3	6.585	5.947	8.121	8.561	7.8705	7.647	7.322	8.451
4	m4	0.0265	0.028	0.029	0.033	0.026	0.025	0.024	0.026
5	m5	0.0975	0.085	0.058	0.066	0.095	0.102	0.068	0.066
6	m6	0.049	0.048	0.037	0.044	0.0405	0.083	0.035	0.038
7	m7	0.0835	0.069	0.066	0.07	0.061	0.072	0.057	0.047
8	m8	0.3705	0.29	0.332	0.396	0.3645	0.274	0.265	0.301
9	m9	1.2	1.308	1.104	1.167	1.5505	0.659	0.79	1.25
10	m10	2.311	2.101	2.128	2.243	2.2445	2.047	2.016	2.533
11	m11	0.0115	0.01	0.012	0.009	0.0125	0.012	0.011	0.01
12	m12	1.415	1.308	1.347	1.024	1.436	1.452	1.449	1.027
13	m13	1.564	1.455	1.507	1.156	1.576	1.588	1.584	1.146
14	m14	0.0235	0.023	0.023	0.018	0.023	0.022	0.023	0.016

図 1、例データ（”Mdata.txt”）

以下では、このデータ(”Mdata.txt”)を用いて、R による階層的クラスタリングの手順を示します。R に入力するコマンドは赤字で示しています。階層クラスタリングでは、用途に応じて距離行列の計算法とクラスター作成法を選ぶ必要がある。トランスクリプトーム解析では距離の計算に(ピアソンの)相関行列を用いることが多く、クラスターは最遠隣法(complete 法)を用いることが多い。

```
#データの読み込み
data <- read.table("Mdata.txt", header=T)

#1-相関係数を距離とする最遠隣法(complete)クラスター解析
c1 <- hclust(as.dist(1-cor(t(data))), method = "complete") #行のクラスタリング
c2 <- hclust(as.dist(1-cor(data)), method = "complete")  #列のクラスタリング
plot(c1, hang=-1)  #行データのクラスタリング結果を描画 hang=-1 で脚を揃える
plot(c2, hang=-1)  #列データのクラスタリング結果を描画 hang=-1 で脚を揃える
```

#ヒートマップ用に距離行列を整える

```
correlationDistanceTable <- as.dist(1-cor(as.matrix((data)))) # 距離行列用
```

#ヒートマップを pdf に出力する　file 名は適宜変更する

```
pdf(file="heatmap.pdf") # pdf に出力する height=10, width=10 などとしてサイズを変更可
  heatmap(x=as.matrix(data.z),Colv=as.dendrogram(c2),Rowv=as.dendrogram(c1),
  distfun=correlationDistanceTable, hclustfun=function(x)
  hclust(correlationDistanceTable),cexRow=0.55, margins = c(6,10))
dev.off()
```

・ユークリッド距離で解析する場合、行データの as.dist(1-cor(t(data)))を dist(data)に、列データの as.dist(1-cor(data))を dist(t(data))に変更する。

・hclust 関数の method 指定で、最近隣法は single、群平均法(UPGMA ともいう)は average、ウォード法は ward などに変更できる。

・cexRow は行ラベルのフォントサイズで、0.2+1 / log10(行数)を基準に変更する。

・margins=(列、行)は、列や行ラベルの文字列の長さに応じて増やす。

以上の操作をまとめると、

```
data <- read.table("Mdata.txt", header=T)
c1 <- hclust(as.dist(1-cor(t(data))), method = "complete")
c2 <- hclust(as.dist(1-cor(data)), method = "complete")
correlationDistanceTable <- as.dist(1-cor(as.matrix((data))))
pdf(file="heatmap.pdf")
  heatmap(x=as.matrix(data),Colv=as.dendrogram(c2),Rowv=as.dendrogram(c1),
  distfun=correlationDistanceTable, hclustfun=function(x)
  hclust(correlationDistanceTable), cexRow=0.55, margins = c(6,10))
dev.off()
```