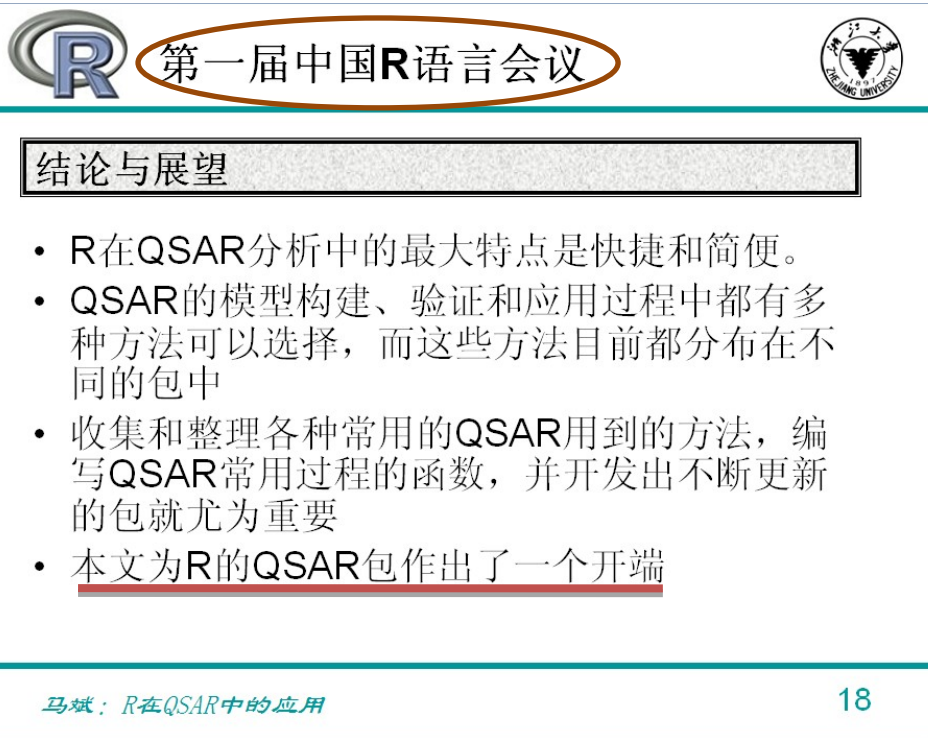
omd 的由来与 QSAR

马斌 徐建明
浙江大学环境与资源学院


第一届中国R语言会议
结论与展望
• R在QSAR分析中的最大特点是快捷和简便。
• QSAR的模型构建、验证和应用过程中都有多种方法可以选择，而这些方法目前都分布在不同的包中
• 收集和整理各种常用的QSAR用到的方法，编写QSAR常用过程的函数，并开发出不断更新的包就尤为重要
• 本文为R的QSAR包作出了一个开端
马斌：R在QSAR中的应用
18

第二届中国 R 语言会议 (北京)

结论与展望

- R在QSAR分析中的最大特点是快捷和简便。
- QSAR的模型构建、验证和应用过程中都有多种方法可以选择，而这些方法目前都分布在不同的包中
- 收集和整理各种常用的QSAR用到的方法，编写QSAR常用过程的函数，并开发出不的包就尤为重要
- 本文为R的QSAR包作出了一个开端

第二届中国 R 语言会议 (北京)

omd: filter the molecular descriptors for QSAR

This package including two useful function, which can be used for filter the molecular descriptors matrix for QSAR.

Version: 1.0
Published: 2009-11-03
Author: Bin Ma
Maintainer: Bin Ma <binma01 at gmail.com>
License: GPL (≥ 2)
CRAN checks: omd results

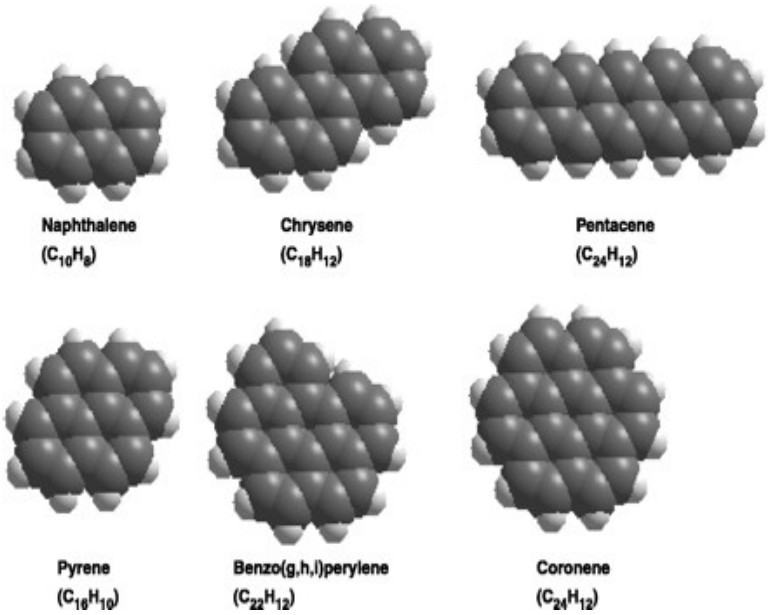
Downloads:

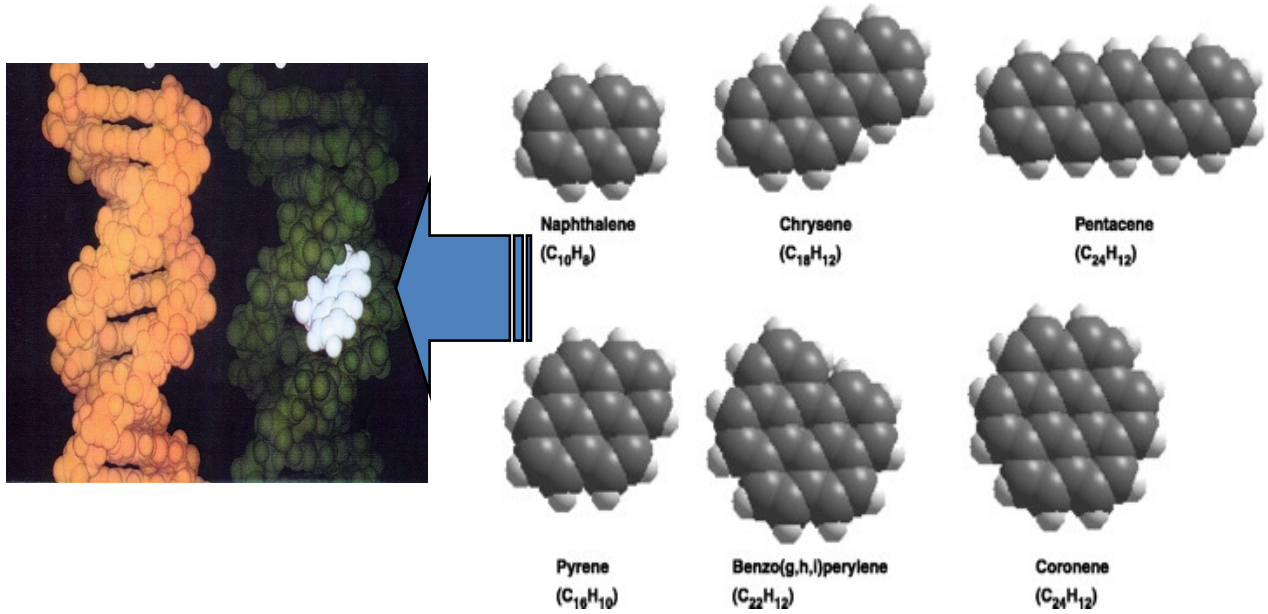
Package source: omd_1.0.tar.gz
MacOS X binary: omd_1.0.tgz
Windows binary: omd_1.0.zip
Reference manual: omd.pdf

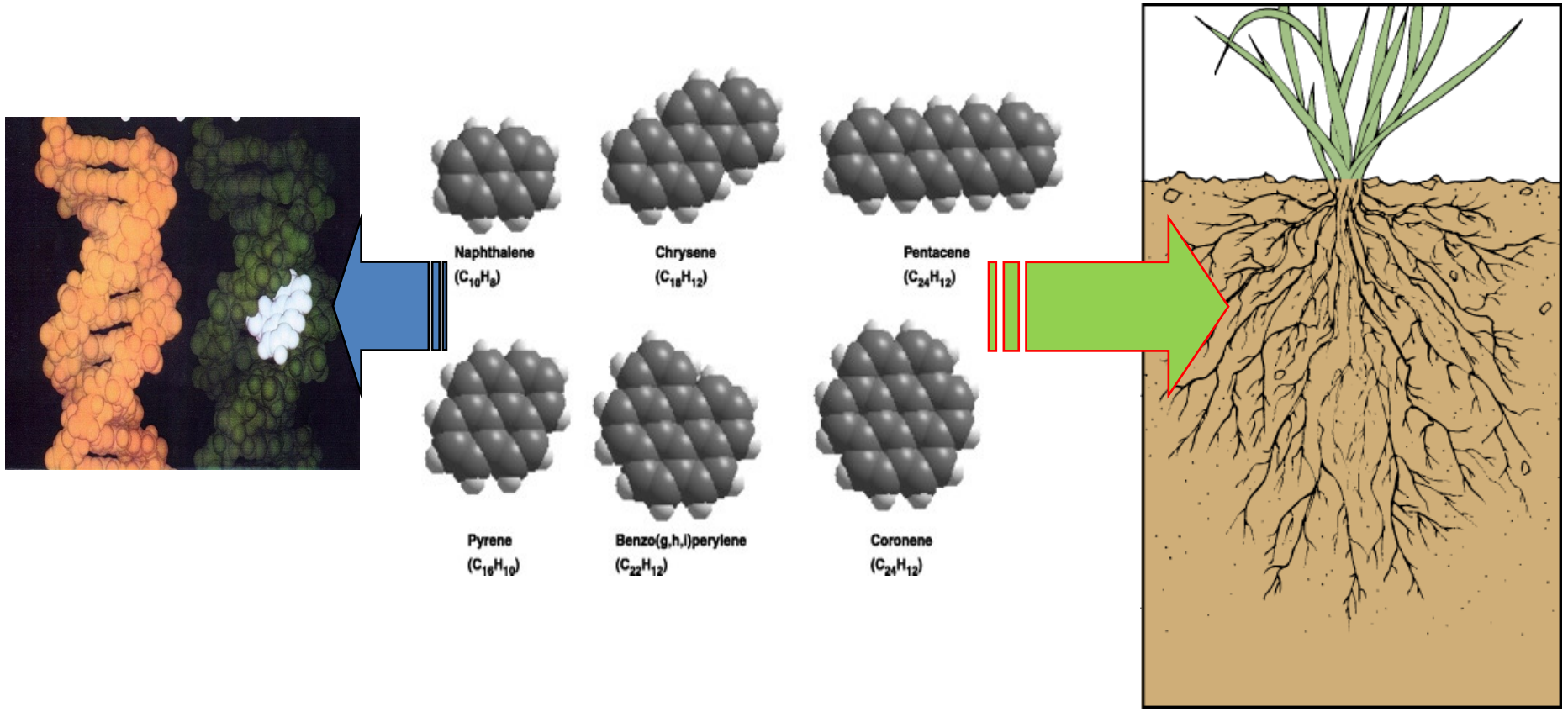
Thanks to *Yihui Xie* from COS and *Kurt.Hornik* from R-group

http://cran.r-project.org/web/packages/omd/index.html

- QSAR （ Quantitative Structure-Activity Relationship ）——**定量构效关系**
- 化合物结构与其效应之间的定量关系，即借助结构参数构建数学模型来描述化合物结构与活性之间的关系
- 活性——化合物的反应活性，比如药效、反应速度、吸附特性……


Naphthalene
($C_{10}H_8$)
Chrysene
($C_{18}H_{12}$)
Pentacene
($C_{24}H_{12}$)
Pyrene
($C_{16}H_{10}$)
Benzo(g,h,i)perylene
($C_{22}H_{12}$)
Coronene
($C_{24}H_{12}$)


Naphthalene
($C_{10}H_8$)
Chrysene
($C_{18}H_{12}$)
Pentacene
($C_{24}H_{12}$)
Pyrene
($C_{16}H_{10}$)
Benzo(g,h,i)perylene
($C_{22}H_{12}$)
Coronene
($C_{24}H_{12}$)


Naphthalene
($C_{10}H_8$)
Chrysene
($C_{18}H_{12}$)
Pentacene
($C_{24}H_{12}$)
Pyrene
($C_{16}H_{10}$)
Benzo(g,h,i)perylene
($C_{22}H_{12}$)
Coronene
($C_{24}H_{12}$)

- **fcons**: filter the constant and nearly constant indices
- **fhcor**: filter the highly correlated indices

fcons

```
# filter the constant and nearly constant indices
fcons<-function(indices,k){
samp<-c()
dim.ind<-dim(indices)
for(i in 1:(dim.ind[2])){
 if (length(unique(indices[,i]))<=k)
 samp<-c(samp,i)
}
indices.fil<-indices[,-samp]
indices.fil
}
```

- fcons 可以将 QSAR 中的分子结构参数矩阵中衡量或者近似衡量的参数删除。
- 通过 Dragon 5.0 计算的 16 种 PAHs 分子结构参数矩阵为 16×1441 。
- 使用 fcons 函数去掉取值少于 3 个的分子结构参数，矩阵简化为 16×226 。

fhcor

```
#filter the highly correlated indices

fhcor<-function(indices,k){
cor.matrix<-cor(indices)
dim.cor<-dim(cor.matrix)
samp<-c()
for(i in 1:(dim.cor[1]-1)){
 for(j in (i+1):dim.cor[2]){
  if(abs(cor.matrix[i,j])>=k){
  samp<-c(samp,j);
  break}
}}
indices.fil<-indices[,-samp]
}
```

- fhcor 可以将分子结构参数矩阵中相关系数高于设定值的参数删除。
- 利用 fhcor 将分子结构参数矩阵简化为 16×48 。
- 如果分子结构参数矩阵仍需要优化，可以采用遗传算法进一步优化。

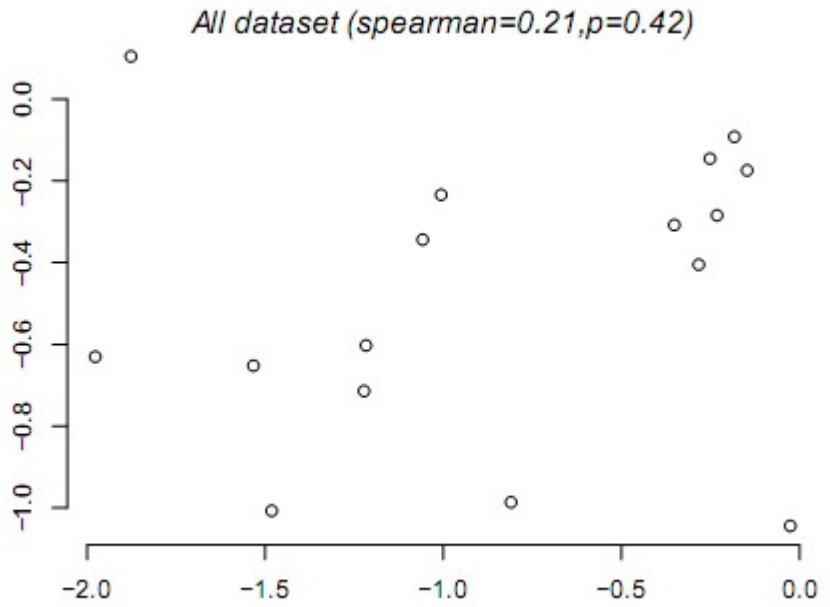
Genetic Algorithm –Partial Least square

```
#GA-PLS

library(genalg)
evalVals<-function(chromosome=c()) {
    returnVal = 1
    Y<-c()
    if (sum(chromosome)>2) {
    MOL = mol3[,chromosome==1];
    PLS<-plsr(lnR~.,data=MOL,method='simpls',validation='CV');
    rmsep<-RMSEP(PLS)
    returnVal <-min(rmsep$val[1,1,])
}
returnVal
}

rbga.results.pls4 = rbga.bin(size=48, zeroToOneRatio=5,
evalFunc=evalVals, popSize=200, iters=100,verbose=TRUE)
```

- 通过 GA-PLS ，分子结构参数矩阵简化为 16×14 。
- 构建优化的分子结构参数矩阵与 PAHs 根际降解效应值之间的 PLS 模型。



The relationship between predicted and calculated effect sizes

Conclusion

- The promised package *omd* have been released online.
- I have grow up from a **green hand** to a R-user putting a foot in the door.

第二届中国 R 语言会议 (北京)

- 眼睛一睁一闭，两天就过去了
- 北京分会就要胜利闭幕了
- 眼睛再一睁一闭，一周就过去了
- 上海分会又要隆重开幕了
- 眼睛再一睁一闭，一年就过去了
- 第三届中国 R 语言会议估计又要进入不紧张的筹备阶段了

- 感谢***第二届中国 R 语言会议会务组***出色的组织