- 2021-06-22
  - GRE 数学部分
    - Problems:
  - 完成安装Docker metaboAnalystic4.0 into server 42.193.18.116
    - 注意事项
    - Probelmes：
    - 开启服务器命令
  - HMMGnen
    - 状态
- 2021-6-23
  - GRE 数学部分
    - Problems
  - GeneHMM 学习 包括隐马尔可夫
    - 复习HMM三类问题
    - 生物序列对比的HMM变种
    - GLIMMER-HMM Based on GHMM
- 2021-6-24
  - MetaboAnalyst 4.0 Tutorial
    - Statical Analysis
    - 代谢组进展
    - 待进展
  - GRE机经
    - 地球月球撞击文章注意点
    - 14修正案
    - 法国革命2月
  - GLIMMER-HMM 原理
- 2021-6-25
  - GRE数学
- 2021-6-26
  - GRE 数学
- 2021-6-28
  - PLAN
  - 代谢组结果进展
    - 分析文件级标准操作流程
    - 分组对应
    - 12h处理组
    - 72h处理组
  - 宏基因组处理

        - 宏基因组分析流程
        - 分箱解析
        - MetaWRAP
        - 运行三种分箱软件
        - Bin提纯
        - Bin注释
    - GLIMMER-HMM
        - 文献阅读
- 2021-6-29
    - PLAN
    - Gene Prediction总结
    - Homology method
        - ETS 方法
        - DP method
        - HMM method
- 2021-6-30
    - PLAN
    - 桉树宏基因组功能丰度可视化

2021-06-22

GRE 数学部分

Problems:

- reciprocal 倒数
- rhombus 菱形
- face card JQK 牌

完成安装Docker metaboAnalystic4.0 into server 42.193.18.116

注意事项

命令为

```
docker pull jsychong/metaboanalyst:august2020
```

Probelmes：

- 可能遇到的问题为服务器体量不够需要扩展服务器。
- 目前有番茄的代谢组，利用这些数据进行测试。

开启服务器命令

```
$ sudo su #进入终端
$ docker run -ti --rm --name metaboanalyst_docker -p 8080:8080 jsychong/metaboanalyst
root@760b678fd4bf:/ Rscript /metab4script.R
root@:/ java -jar /opt/payara/payara-micro.jar --deploymentDir /opt/payara/deployments #这种情况
```

这种情况下关闭服务器网站依然运行可以进行运行

网站网址 http://42.193.18.116:8080/MetaboAnalyst/

HMMGnen

状态

4种状态 E1 E2 E3 I
Ei 碱基位置为第i号密码子 i=1,2,3

- Profile HMM gene prediction
- CONTEXT-SENSITIVE HMMS AND PROFILE-CSHMMS
- 基因预测和序列对比，内含子外显子识别的对比利用的HMM

2021-6-23

GRE 数学部分

Problems

- For which of the following values of x is the units digit of the product $2 * 3^x$ equal to 4?

解题思路： 3的平方digital unit规律 3 9 27

- prependicular 垂直于 用于计算k line
- parallel counterpart类似物

GeneHMM 学习 包括隐马尔可夫

复习HMM三类问题

- 给定Obs 求概率

Notes: 利用动态规划方法求解 向前方法，向后方法，公式推导--全概率公式

- 给定Obs 求解隐藏状态路径

Notes: Vertebi Algorithm 与向前方法类似但记录的值为上一个状态的后验概率最大值。向前向后算法，不仅可以输出路径，还可以输出每一个个体的可靠性

- 给定Obs 求参数

Notes: 首先随机生成隐藏序列，估计参数，再反推隐藏序列，再估计参数，直至收敛

生物序列对比的HMM变种

- GHMM Generalized HMM

普通的HMM根据一步转移概率进行推算，时间间隔是指数分布，显然在解决问题不合适

特点：每一个状态到另一个状态的时间间隔Distribution 不同 state - duration(间隔长度) s_1对应间隔时间的概率密度函数 $f_{s_1}(d)$

应用：多用于基因预测

- PHMM Paired HMM

两个output序列,共享一个HMM参数，即相同的状态空间，相同状态对应相同转移概率发射概率

e.g. Sequence S1 & Sequence S2 For a shared PHMM each base of both of the two sequence have the same hidden statues such as matched, mismatched, gap...

- GPHMM

有以上两个变种的特征，但记录单位改变为每个状态+间隔

e.g. for two sequences we denote (d,e) for the duration tuple of the two sequences. For the first segment maybe the statues is s_i the duration tuple is (d_1, e_1). It means from 0 to d_1 nt or base the segment's status is s_i in sequence S1 and from 0 to e_1 nt or base the segment's status is s_i in sequence S2

GLIMMER-HMM Based on GHMM

- Core Algorithm

$$\underset{\phi}{argmax} \prod_{q \in \phi} P(S_i|q_i, d_i)P(q_i|q_{i-i})P(d_i)$$

S denotes sequence
q_i denotes statues
ϕ denotes the optimium path

2021-6-24

MetaboAnalyst 4.0 Tutorial

Statical Analysis

可以进行PCA PLS-DA分析聚类，如果是2-factor可以绘制火山图，fold-change图 主要通过这个功能进行操作 最重要注意的是数据清洗
使用方法：

- 服务器网站 http://42.193.18.116:8080/MetaboAnalyst
- 在使用ssh登陆时修改命令

```
$ ssh -o ServerAliveInterval=30 @42.193.18.116
```

代谢组进展

- 自然处理、超促排卵、羟基脲超促排卵4度和25度处理的PLS-DA进行清晰分离

待进展

- 探究区分方法数学原理
- 提出预测指标RIC SVM预测并自学SVM

GRE机经

地球月球撞击文章注意点

- titantic撞击可能会过于有吸引力，所以很多现象被错误的解释了
- component and elements 问题，可能在地球上但没再约六上 *原因是abensence in lunar rock*

14修正案

- 主旨 二战后最高法院同意公民权
- state action 的原因：一定程度保留歧视
- 修正案最初支持者的原因是在于 支持人权平等

法国革命2月

- author agree 不同的革命作用不同
- 对最后一段的反驳 1830
- 第二段最主要说明description的重要

GLIMMER-HMM 原理

- 完成了编译运行 tips:将c代码转化为cpp才能通过编译

2021-6-25

GRE数学

> The width and the length of a rectangular piece of plywood are 4 feet and 8 feet, respectively. Along one edge of the plywood, a strip x inches wide and 8 feet long is removed. Then, along an edge perpendicular to the 8-foot edge, a strip x inches wide is removed. For what value of x will the remaining rectangular piece have width and length in the ratio of 2 to 5? (1 foot = 12 inches)

解析：long one edge of the plywood, a strip x inches wide and 8 feet long is removed. 长8 feet 宽x inch的长方形直板被移除

实际问题： **将长方形纸板去掉x inches后长宽比例5/2 求x值**

2021-6-26

GRE 数学

Revenue 销量
Isosceles triangle 等腰三角形

2021-6-28

PLAN

- **代谢组数据分析并制作ppt**
- **夏令营推荐信签字盖章**
- **宏基因组开头指定研究计划总结研究方法**
- **GRE机经下载**
- GLIMMER-HMM 文件格式探究

代谢组结果进展

分析文件级标准操作流程

- 选定统计分析 网址为 http://42.193.18.116:8080/MetaboAnalyst/faces/upload/StatUploadView.xhtml
- 选定文件 Data_Pos_qc - 副本.csv 即已经整理好的代谢信息 并修改表头分组表示时间12&72

分组对应

- ZR 自然
- CP 超促排卵
- QJ 羟基脲25
- QJCP 羟基脲25+超促排卵
- YH 羟基脲4
- YHCP 羟基脲25

- ZC WS ZCWS 羟基脲25+超促排卵+中药

12h处理组

绘制四图，PCA-PLS-DA图、代谢组特征相关性图、热土、SAM 图以及筛选和FDR值0.000112 δ值为1.4 *SAM:多组间表达量差异分析*
不需要剔除数据即可

72h处理组

绘制四图，PCA-PLS-DA图、代谢组特征相关性图、热土、SAM 图以及筛选和FDR值 *SAM:多组间表达量差异分析*
需要剔除数据 YHCP72-4 ZR72-1

宏基因组处理

特点

- 物种丰度 可选 扩展到种，已有分析好数据基于K数据库
- 基因预测和聚类 KEGG CAZY
- 分箱 Binning

宏基因组分析流程

- 打碎DNA，根据相似性拼接
- 识别基因，预测基因
- 定量计算基因丰度
- 翻译成蛋白，和数据库对比 --计算功能丰度
- 分箱
- 作图网站 www.ehbio.com/ImageGP

分箱解析

Binning的含义是分箱、聚类，指从微生物群体序列中将不同个体的序列（reads或contigs等）分离开来的过程。简单来说就是把宏基因组数据中来自同一菌株的序列聚到一起，得到一个菌株的基因组。是的，可以达到菌株水平。

reads Binding -- Contig binning

利用序列进行分箱

MetaWRAP

```
# 主要使用MetaWRAP，演示基于官方测试数据
# 主页：https://github.com/bxlab/metaWRAP
# 挖掘单菌基因组，需要研究对象复杂度越低、测序深度越大，结果质量越好。要求单样本6GB+，复杂样本如土壤推荐
# 上面的演示数据12个样仅140MB，无法获得单菌基因组，这里使用官方测序数据演示讲解
# 软件和数据库布置需2-3天，演示数据分析过程超10h，标准30G样也需3-30天，由服务器性能决定。
```

准备数据和环境变量

```
# 流程：https://github.com/bxlab/metaWRAP/blob/master/Usage_tutorial.md
>
# 输入数据：质控后的FASTQ序列，文件名格式必须为*_1.fastq和*_2.fastq
#           C1_1_kneaddata_paired_1.fastq  -> C1_1_1.fq
#           C1_1_kneaddata_paired_2.fastq  -> C1_1_2.fq
#           放置到 binning/temp/qc 目录下

# 拼装获得的contig文件：result/megahit/final.contigs.fa
#           放置到 binning/temp/megahit 目录下
#

# 中间输出文件：
#     Binning结果：binning/temp/binning
#     提纯后的Bin统计结果：binning/temp/bin_refinement/metawrap_50_10_bins.stats
#     Bin定量结果文件：binning/temp/bin_quant/bin_abundance_heatmap.png
#                      binning/temp/bin_quant/bin_abundance_table.tab (数据表)
#     Bin物种注释结果：binning/temp/bin_classify/bin_taxonomy.tab
#     Prokka基因预测结果：binning/temp/bin_annotate/prokka_out/bin.10.ffn 核酸序列
#     Bin可视化结果：binning/temp/bloblogy/final.contigs.binned.blobplot (数据表)
#                    binning/temp/bloblogy/blobplot_figures (可视化图)

# 准备原始数据从头分析，详见公众号或官网教程
# 这里我们从质控后数据和拼接结果开始
cd ${wd}
mkdir -p binning && cd binning
mkdir -p temp && cd temp
# 这里基于质控clean数据和拼接好的contigs，自己链接自上游分析
# 7G质控数据，输入数据文件名格式必须为*_1.fastq和*_2.fastq
mkdir -p seq
cd seq
# 方法1. 下载测序数据
# for i in `seq 7 9`;do
#    wget -c http://210.75.224.110/share/meta/metawrap/ERR01134${i}_1.fastq.gz
#    wget -c http://210.75.224.110/share/meta/metawrap/ERR01134${i}_2.fastq.gz
# done
# gunzip *.gz # 解压文件
# rename .fq .fastq *.fq # 批量修改扩展名
# 方法2. 复制准备好的数据
ln -sf ${db}/metawrap/*.fastq ./
cd ..
# megahit拼接结果
mkdir -p megahit
cd megahit
# wget -c http://210.75.224.110/share/meta/metawrap/final.contigs.fa.gz
# gunzip *.gz
ln -s ${db}/metawrap/*.fa ./
cd ../..

# 加载运行环境
cd ${wd}/binning
conda activate metawrap
```

运行三种分箱软件

```
metawrap -v
# 输入文件为contig和clean reads
# 调用三大主流binning程序cococt，maxbin2，metabat2
# 8p线程2h，24p耗时1h
# nohup 和 & 保证任务在后台不被中断，且记录输出内容到 nohup.out(可选)
nohup metawrap binning -o temp/binning -t 1 -a temp/megahit/final.contigs.fa \
  --metabat2 --maxbin2 --concoct temp/seq/ERR*.fastq &
# 用自己的文件，替换输出文件名为 *1_kneaddata_paired*.fastq
# 如果想接上上面的流程使用自己的文件做分析，则把ERR*.fastq替换为 *1_kneaddata_paired*.fastq
# 输出文件夹 temp/binning 包括3种软件结果和中间文件
```

Bin提纯

```
# 8线程2h， 24p 1h
cd ${wd}/binning
# rm -rf temp/bin_refinement
metawrap bin_refinement \
  -o temp/bin_refinement \
  -A temp/binning/metabat2_bins/ \
  -B temp/binning/maxbin2_bins/ \
  -C temp/binning/concoct_bins/ \
  -c 50 -x 10 -t 2
# 查看高质量Bin的数量，10个，见temp/bin_refinement/metawrap_50_10_bins.stats目录
wc -l temp/bin_refinement/metawrap_50_10_bins.stats
# 结果改进程度见temp/bin_refinement/figures/目录
```

Bin注释

```
# Taxator-tk对每条contig物种注释，再估计bin整体的物种，11m (用时66 min)
metawrap classify_bins -b temp/bin_refinement/metawrap_50_10_bins \
  -o temp/bin_classify -t 2 &
# 注释结果见`temp/bin_classify/bin_taxonomy.tab`

# export LD_LIBRARY_PATH=/conda2/envs/metagenome_env/lib/:${LD_LIBRARY_PATH}
 # 这是动态链接库找不到时的一个简单的应急策略
ln -s /conda2/envs/metagenome_env/lib/libssl.so.1.0.0 .
ln -s /conda2/envs/metagenome_env/lib/libcrypto.so.1.0.0 .

# 基于prokka基因注释，4m
metaWRAP annotate_bins -o temp/bin_annotate \
  -b temp/bin_refinement/metawrap_50_10_bins  -t 1
# 每个bin基因注释的gff文件bin_funct_annotations,
# 核酸ffn文件bin_untranslated_genes,
# 蛋白faa文件bin_translated_genes
```

GLIMMER-HMM

文献阅读

两种方法 DP方法 Markov 模型

2021-6-29

PLAN

- **完成文献阅读(Computational gene finding in plants)**
- **检查SaAlign论文并对应图片**
- *检查根际微生物论文文字和对应关系*
- *GRE 填空3section 阅读3section*

Gene Prediction总结

Gene prediction methods are based on siganl or homology.
For different types of genome we have different states like when we are dealing with prokarytoes, we only focus on identifying the start codons and stop condons. For Eukaryotes, except for start codons and stop condons, exons and introns are oftern taken into consideration.
The core idea of methods for signals is if x is a type of signal $length(x) = l$,

$$P(x) = \prod_{i=1}^{l} p^{(i)}(x_i)$$

Gene prediction methods are based either DP method or HMM method.

Homology method

ETS 方法

Genes can also be identified by homology with expressed sequence tags (ESTs) (Franco et al., 1995). EST data are usually generated from single-pass sequences and are therefore less accurate than genomicdata, but when an EST matches a gene, it can givevery precise gene location information.

DP method

A wrapper method algorithm conto combine all the possible signals.

> **Application**
>
> 1. Find all the puative 假定的 exons
> 2. Then find the most possible path

HMM method

Core Algorithm

$$logP(S) = \sum_{i=1}^{n_k} logP(b_i|b_{i-1}, b_{i-2}...b_{i-5})$$

5-th Markov model

- Output state: ATCG....
- Hidden state: exon. intron, intergenetic region

GHMM

> introduce a duration time of every hidden state
> So we can determine the duration time distribution

IMM

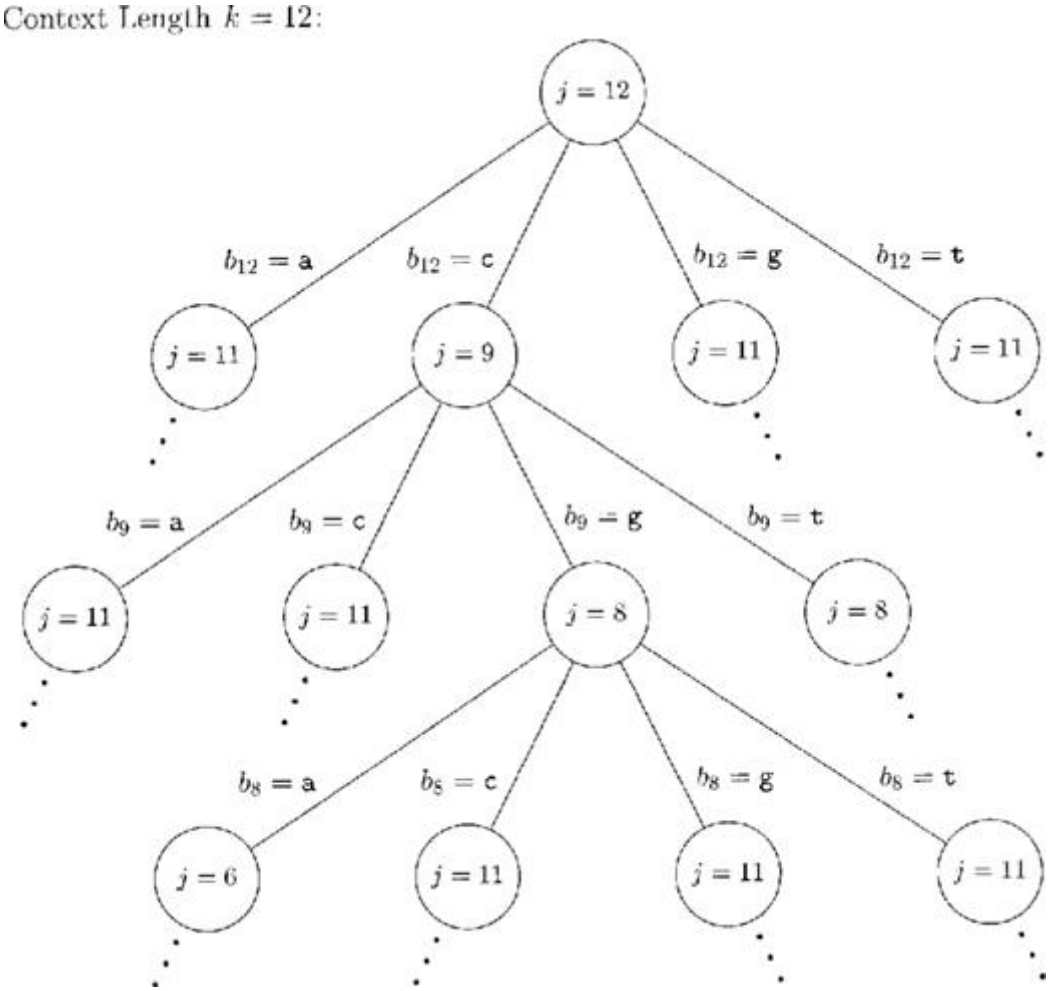
> focus on the problem of high-order e.g. for 5-th order we need to record $4^5 = 4096$ records (transmission probability)
> Core formula
>
> $$IMM_k(S_x) = \lambda_k(S_{x-1}) * P_k(S_{x-1}) + [1 - \lambda_k(S_{x-1})] * IMM_{k-1}(S_x)$$
>
> IMM denotes the score of IMM analysis S_k

ICM

与IMM的区别在于基于context的ICM首先需要判断交互信息值the mutual information values

$$I(X;Y) = \sum_i \sum jP(x_i, y_j) log(\frac{P(x_i)P(y_j)}{P(x_i, y_j)})$$

Then find the maximium and set ut as the root node like the graph. then set b_k as a,t,c and g then get the min I(x_k,x_j) then until the bottom of the tree. Then use the method of IMM. IMM 是顺序递归、即IMM_k的值与IMM_{k-1}有关。对于ICM 则从底部上升如图IMM_9的值与IMM_8有关IMM_{12}的值与IMM_9有关



2021-6-30

PLAN

- **SaAlign 文章投出**
- VirFinder 文献阅读
- **检查根际微生物论文文字和对应关系**
- **桉树宏基因组功能丰度可视化**
- **GRE高高频3+3**

桉树宏基因组功能丰度可视化

主要针对GO数据和KEGG数据寻找与氮代谢相关的GO或者KO
首先判断差异表达量 即ORG VS Y5 & Y5 VS Y10 在夏天和冬天的改变&并做图

- KEGG 功能丰度热图和检验