PSMC 测试说明文档

背景

溯祖模拟是一种对群体遗传学信息根据时间向后推算的模型（backward model）。对祖先的模拟是通过寻找到最近共同祖先（Most Recent Common Ancestor，MRCA）完成。

关于现代人类的起源一直存在着争议,但目前绝大多数学者接受“走出非洲”学说,即认为:十万年前现代人起源于非洲的某个较小人群,随着人口数量的不断增长,逐步迁移到非洲以外的很多地方。在过去近三十年里,遗传学不断地渗透到人类起源的研究中,并且提供了众多支持“走出非洲”学说的 DNA 分子水平证据。在此基础上,有学者进一步提出了“系列建立者模型”学说:当撒哈拉以南非洲某个群体达到一定人口饱和度时,其中的一部分人就会迁移到远离其原居住地的另外一个地区建立一个新群体,当新群体的人口再次增长到较高的饱和度时,就会发生第二次迁移,这个过程反复不断,直至达到地球上最远的人类聚居点——美洲大陆。本研究应用遗传学数据和遗传统计学原理系统地探讨了这个过程中的人口变迁历史,如世界不同地区群体的最近共同祖先年代(TMRCA)、有效群体的大小(Ne)、人口数量增长的起始时间以及人口的增长率等问题

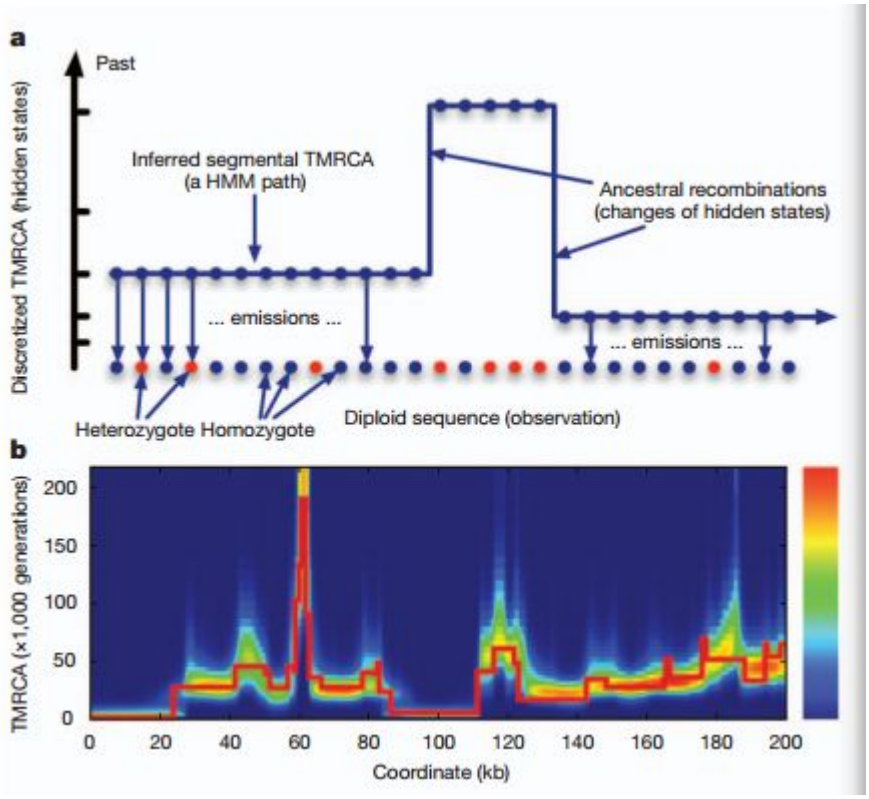
理论

PSMC(Pairewise Sequentially Markovian Coalescent),成对马尔科夫溯祖模型是根据一个二倍体基因组数据模拟出祖先有效群体大小的溯祖模型，它使用了隐马尔科夫链模型，把历史祖先看做是一个隐含未知参数的马尔科夫过程，从一个个体的全基因组数据这个可观察的参数确定整个过程的隐含量。

模型的解释可以参考下图：

一个二倍体的个体含有几十万，上百万的独立的等位基因点，每个位点的两个等位基因都可以找到其最近的公共祖先（TMRCA），如下图 a 所示，进而 TMRCA 的分布情况可以由每段等位基因的状态进行推导得到：整个 genome 上等位基因杂合程度的分布可以反映出由于重组时间导致 TMRCA 的分离（a 图中 TMRCA 的状态）

下图 b 中：仿真了 200k 的染色体上 TMRCA 的分布情况



主要理论模型如下所示：

1.连续 PSMC 模型

把观测到的基因组数据和 reference 比对，提取 consensus 文件，"1"表示纯合，"0"表示杂合，"." 表示未知，那么 HMM 中从状态 t 的 emission probability:

$$e(1\,|\,t)=e^{-\theta t}$$

$$e(0\,|\,t)=1-e^{-\theta t}$$ 其中 θ 表示突变率

$$e(.\,|\,t)=1$$

从 s 状态转换成 t 状态的 transition probability:

$$p(t\,|\,s)=\Pr\{T_{a+1}=t\,|\,T_a=s\}=(1-e^{-\rho t})q(t\,|\,s)+e^{-\rho t}\delta(t-s)$$

$$q(t\,|\,s)=\Pr\{T_{a+1}=t\,|\,T_a=s,R_a=1\}=\frac{1}{\lambda(t)}\int_0^{\min\{s,t\}}\frac{1}{s}\times e^{-\int_u^t\frac{dv}{\lambda(v)}}du$$

$$\sigma(t)=\Pr\{T_a=t\}=\frac{\pi(t)}{C_\sigma(1-e^{-\rho t})}$$

$$\pi(t)=\Pr\{T_{a+1}=t,R_a=1\}=\frac{t}{C_\pi\lambda(t)}e^{-\int_0^t\frac{dv}{\lambda(v)}}$$

其中 ρ 表示重组率，Ra=1 表示有重组发生　δ(.)是 Dirac delta 函数

$\lambda(t)=N_e(t)/N_0$ 是指在 t 状态下的相对群体大小

2 离散 PSMC 模型

把溯祖时间分成很多间隔以及在每个间隔里面的 emission probability 和 transition probability。

$$t_i = 0.1(e^{\frac{i}{n}\log(1+10T_{\max})} - 1)$$

在时间间隔$[t_k, t_{k+1})$， λ（t）可以看做常数λ_k

$$\pi_\kappa = \int_{t_k}^{t_{k+1}} \pi(t)dt$$

$$\sigma_\kappa = \int_{t_k}^{t_{k+1}} \sigma(t)dt$$

$$q_{kl} = \frac{1}{\pi_\kappa} \int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} q(t \mid s)\pi(s)dt$$

$$p_{kl} = \frac{1}{\sigma_\kappa} \int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} p(t \mid s)\sigma(s)dt$$

$$\pi_\kappa = \frac{1}{C_\pi}[(\alpha_\kappa - \alpha_{\kappa+1})(\sum_{i=0}^{k-1} \tau_i + \lambda_\kappa) - \alpha_{\kappa+1}\tau_k]$$

$$\sigma_\kappa = \frac{1}{C_\rho}[\frac{1}{C_\pi \rho}(\alpha_\kappa - \alpha_{\kappa+1}) + \frac{\pi_\kappa}{2} + o$$

其中：

$$\tau_k = t_{\kappa+1} - t_\kappa$$

$$\alpha_\kappa = \exp(-\sum_{i=0}^{k-1} \frac{t_{i+1} - t_i}{\lambda_i})$$

$$C_\pi = \sum_{k=0}^{n} \lambda_\kappa(\alpha_\kappa - \alpha_{\kappa+1})$$

测试

数据来源

ftp://ftp.sanger.ac.uk/pub/rd/humanSequences/

“Inference of human population history from individual whole-genome sequences”中使用的一个NA18507.fq.gz 常染色体数据 与 reference 进行过比对的 consensus sequence 数据

注意： 如果是下机的数据需要经过如下步骤得到 consensus sequence 文件： 1.bwa 比对得到 bam 文件 2.经过 samtools 处理得到 fq 文件

```
samtools mpileup -C50 -uf ref.fa aln.bam | bcftools view -c - \
        | vcfutils.pl vcf2fq -d 10 -D 100 | gzip > diploid.fq.gz
```

数据路径：
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/NA18507.fq.gz

测试程序

测试程序：/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc_test.sh

```
### convert fq to psmcfa ,so psmc can interpret
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/fq2psmcfa -q20 NA18507.fq.gz > NA18507.psmcfa

### run psmc
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o NA18507.psmc
NA18507.psmcfa

### history
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/psmc2history.pl NA18507.psmc |
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/history2ms.pl > ms-cmd.s
h
### plot
##add the path for epstopdf
export PATH=/nfs3/onegene/user/group1/guolihua/soft/bin:$PATH
export PATH=/nfs/biosoft/latex/2014/bin/x86_64-linux/:$PATH
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/psmc_plot.pl -p NA18507    NA18507.psmc
```

1.数据前期处理

```
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/fq2psmcfa -q20 NA18507.fq.gz > NA18507.psmcfa
```

输入：与 reference 比对好的 consensus sequence 的压缩文件 NA18507.fq.gz
参数-q20 质量值过滤
输出：NA18507.psmcfa
格式：

```
>1
NNNNKNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNTNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTNT
```

```
NNNNNNNNNNNNNNNNNNNNNNNNTTTTNNTTKTNKNTNTNNNNTTTTNTTTTNNNNTNTN
NNTTNNNNN.....
```

说明：

对 NA18507.fq.gz 进行如下处理，每 100bp 为一个 bin

1.>90 的 bases 为低质量，为 N

2.>10 的 bases 为高质量，至少有一个 heterozygous，为 K，否则为 T

2.PSMC 主程序

```
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o NA18507.psmc
NA18507.psmcfa
```

输入：NA18507.psmcfa 即上一步的输出

参数：

-p STR pattern of parameters [4+5*3+4]
-t FLOAT maximum 2N0 coalescent time [15]
-N INT maximum number of iterations [30]
-r FLOAT initial theta/rho ratio [4]
-c FILE CpG counts generated by cntcpg [null]
-o FILE output file [stdout]
-i FILE input parameter file [null]
-T FLOAT initial divergence time; -1 to disable [-1]
-b bootstrap (input be preprocessed with split_psmcfa)
-S simulate sequence
-d perform decoding
-D print full posterior probabilities

输出：/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/NA18507.psmc

输出格式：

每一行都以两个字母开头，作为一行的 mark 表示的含义参见 CC 注释

```
CC
CC        Brief Description of the file format:
CC          CC    comments
CC          MM    useful-messages
CC          RD    round-of-iterations
CC          LL    \log[P(sequence)]
CC          QD    Q-before-opt Q-after-opt
CC          TR    \theta_0 \rho_0
CC          RS    k t_k \lambda_k \pi_k \sum_{l\not=k}A_{kl} A_{kk}
CC          DC    begin end best-k t_k+\Delta_k max-prob
CC
MM        Version: 0.6.4-r49
MM        pattern:4+25*2+4+6, n:63, n_free_lambdas:28
MM        n_iterations:25, skip:1, max_t:15, theta/rho:5
MM        is_decoding:0
MM        n_seqs:30, sum_L:26405277, sum_n:2175214
RD        0
LK        0.000000
QD        0.000000 -> 0.000000
RI        inf
```

```
TR      0.085970        0.017194
MT      15.000000
MM      C_pi: 1.000000, n_recomb: 450141.116741
RS      0       0.000000        1.000000        3684.543201     0.008185        0.000000
RS      1       0.008290        1.000000        3955.976136     0.008788        0.000000
RS      2       0.017266        1.000000        4244.392803     0.009429        0.000000
RS      3       0.026987        1.000000        4550.339853     0.010109        0.000000
RS      4       0.037514        1.000000        4874.283698     0.010828        0.000000
RS      5       0.048914        1.000000        5216.587893     0.011589        0.000000
RS      6       0.061258        1.000000        5577.487234     0.012391        0.000000
RS      7       0.074626        1.000000        5957.058360     0.013234        0.000000
RS      8       0.089102        1.000000        6355.186694     0.014118        0.000000
。。。。
```

说明：p 参数说明

4+25*2+4+6 时间间隔跨越模式，表示第一个参数跨越 4 个时间间隔，接下去的 25 个参数每个跨越 2 个时间间隔，第 27 个参数跨越 4 个时间间隔，第 28 个参数跨越 6 个时间间隔

3.画图程序

```
### plot
/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc-master/utils/psmc_plot.pl -p NA18507    NA18507.psmc
```

输入：把上一步输出的 psmc 文件放到最后作为 perl 的输入

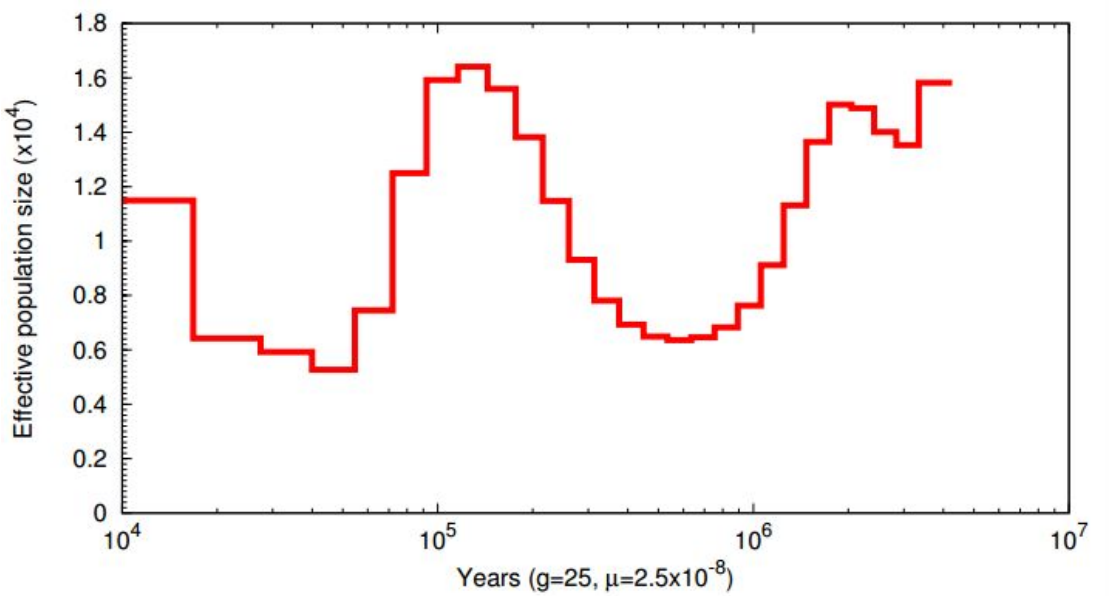
参数：-p 表示最后生成 pdf 格式，由于转换需要用到 epstopdf 程序，所以请保证能找到对应的程序，如下所示

```
export PATH=/nfs/biosoft/latex/2014/bin/x86_64-linux/:$PATH
```

参数说明如下：

```
Options: -u FLOAT     absolute mutation rate per nucleotide [2.5e-08]
         -s INT       skip used in data preparation [100]
         -X FLOAT     maximum generations, 0 for auto [0]
         -x FLOAT     minimum generations, 0 for auto [10000]
         -Y FLOAT     maximum popsize, 0 for auto [0]
         -m INT       minimum number of iteration [5]
         -n INT       take n-th iteration (suppress GOF) [20]
         -M titles    multiline mode [null]
         -f STR       font for title, labels and tics [Helvetica,16]
         -g INT       number of years per generation [25]
         -w INT       line width [4]
         -P STR       position of the keys [right top]
         -T STR       figure title [null]
         -N FLOAT     false negative rate [0]
         -S           no scaling
         -L           show the last bin
         -p           convert to PDF (with epstopdf)
         -R           do not remove temporary files
         -G           plot grid
```

画图：



说明：

1）画图程序默认选取第 20 次迭代的值

2）X 轴为时间，计算方法：2N0*t_k*generation

3）Y 轴为预测群体大小，计算方法：N0* λ_k /10000

其中 N0 = θ / 4 μ

θ 在 psmc 的 RT 一行中给出，μ在画图程序中（u）可以设置, 默认值为 2.5e-8

Bootstrap 验证

Bootstrap 又称为 bootstrap 重复取样，是为了估计样本数据的统计精度，计算每一次 bootstrap 取样的 psmc 预测结果。这些值被用来估计原始样本的预测结果的精度。

bootstrap 取样方法的假设：

1）你的样本能有效的代表样本

2）从样本里面再进行有放回取样，每一次子取样都是独立同分布的，换句话说，它假设子样本和总体的分布相同，但每个样本都是和其他样本独立的。

程序

/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/psmc_test_bootstrap.sh

方法

1）把输入文件分割成 30M 大小的 segments
2）使用 bootstrap 的方法，有放回取样，得到和原始二倍体大小一样的新的二倍体
3）对新的二倍体进行 psmc 群体大小预测

结果

/nfs3/onegene/user/group1/guolihua/Test_soft/PSMC/combined.NA18507.pdf