

# 魔法酶切（Magic Enzyme Cutter）软件开发

作者：毛凤彪

学号：201318008515006

导师：孙中生

中科院.北京生命科学研究院



队长：毛凤彪  
学号：201318008515006  
研究所：中国科学院大学  
邮箱：maofengbiao@gmail.com

## 团队成员：



姓名：蔡万世  
学号：201118008515008  
研究所：中国科学院大学  
邮箱：524573104@qq.com



姓名：文艳玲  
学号：201328016715035  
研究所：中国科学院大学  
邮箱：wyling19@126.com

2013 年 12 月 22 日

## 摘 要

为嵌合酶切技术在第二代技术的充分和合理运用,开发出一个灵活的、全面的、高效的、友好的酶切模拟和评估软件是迫在眉睫。因此,我们将开发一款名为“魔法酶切”(magic enzyme cutter)的软件,旨在完成和实现以下目的:

- (1) 采用 BM 算法大规模高效率地模拟单酶或多酶酶切,得到每一个酶切片段的黏端信息和碱基序列,解决以往酶切工具的基因组大小限制问题。
- (2) 统计和显示酶切片段大小、GC 含量和碱基平衡性,以评估特定长度大小片段的扩增效率。
- (3) 以电泳图的形式展示模拟酶切结果,形象生动的展示酶切效果。
- (4) 计算和统计特定长度范围上的特定序列的比例,为特定碱基或序列靶向测序提供依据和支持。
- (5) 使用最优解算法 PSO 优化酶切的区域范围的选择,使所选区域范围对目标碱基或特征序列测序效能最高。

最终将上面的结果以网页的形式汇总成评估报告,方便研究者和使用者一目了然地知道模拟酶切效果。



## 目 录

|                       |    |
|-----------------------|----|
| 研究背景.....             | 4  |
| 一、限制性内切酶综述。.....      | 4  |
| 二、第二代测序的运用.....       | 5  |
| 三、酶切技术在第二代测序中的运用..... | 7  |
| 研究内容.....             | 10 |
| 研究方法.....             | 12 |
| 1、字符串匹配算法.....        | 12 |
| 2、智能优化算法。.....        | 13 |
| 研究结果（部分）.....         | 15 |
| 1、模拟酶切片段分布.....       | 16 |
| 2、模拟酶切胶图展示。.....      | 16 |

# 研究背景

## 一、限制性内切酶综述。

40 多年前，当人们在对噬菌体的宿主特异性的限制-修饰现象进行研究时，首次发现了限制性内切酶。细菌可以抵御新病毒的入侵，而这种"限制"病毒生存的办法则可归功于细胞内部可摧毁外源 DNA 的限制性内切酶。首批被发现的限制性内切酶包括来源于大肠杆菌的 *EcoR I* 和 *EcoR II*，以及来源于 *Haemophilus influenzae* 的 *Hind II* 和 *Hind III*。这些酶可在特定位点切开 DNA，产生可体外连接的基因片段。研究者很快发现内切酶是研究基因组成、功能及表达非常有用的工具。

当限制性内切酶的应用在上世纪七十年代流传开来的时候，以 NEB 为代表的许多公司开始寻找更多的限制性内切酶。除了某些病毒以外，限制性内切酶只在原核生物中被发现。人们正在从数以千计的细菌及古细菌中寻找新的限制性内切酶。而对已测序的原核基因组数据分析表明，限制性内切酶在原核生物中普遍存在--所有自由生存的细菌和古细菌似乎都能编码限制性内切酶。

限制性内切酶的形式多样，从大小上来说，它们可以小到如 *Pvu II*（157 个氨基酸），也可以比 1250 个氨基酸的 *Cje I* 更大。在已纯化分类的 3000 种限制性内切酶中，已发现了超过 250 种的特异识别序列。其中有 30%是在 NEB 发现的。对具有未知特异识别序列的限制性内切酶的研究发现工作仍在继续。人们从分析细胞提取物的生化角度研究的同时，也采用计算机分析已知的基因组数据，以期有更多的发现。尽管很多新发现的酶的识别序列与已有的重复--即同裂酶，仍然有识别新位点的酶不断被发现。

上世纪 80 年代，NEB 开始克隆并表达限制性内切酶。克隆技术由于将限制性内切酶的表达与原有细胞环境分离开来，避免了原细胞中其它内切酶的污染，从而提高了酶的纯度。此外，克隆技术提高了限制性内切酶的产量，简化了纯化过程，使得生产成本显著降低；克隆的基因很容易进行测序分析，表达出的蛋白也能进行 X 射线结晶分析，这使得我们对于克隆产物更加确定。

限制性内切酶的主要功能是保护细菌不受噬菌体的感染，这一观点已被人们广泛接受。它们作为微生物免疫机制的一部分行使其功能。当一个没有限制性内切酶的细菌被病毒感染时，大部分病毒颗粒都能成功地进行感染。然而一个有限制性内切酶的同种细菌被成功感染的比率显著下降。出现更多的限制性内切酶将会起到多重保护作用；而一个拥有 4 到 5 种各自独立的限制性内切酶将会使细胞坚不可摧。

限制性内切酶常常伴随一到两种修饰酶（甲基化酶）出现。后者的作用是保护细胞自身的 DNA 不被限制性内切酶破坏。修饰酶识别的位点与相应的限制性内切酶相同，但只甲基化每条链中的一个碱基，而不是切开 DNA 链。限制性内切酶识别位点处的甲基基团伸入到双螺旋的大沟中去，阻碍了限制性内切酶的作用。这样，限制性内切酶和它的"搭档"--甲基化酶一起就构成了限制-修饰（R-M）系统。在一些 R-M 系统中，限制性内切酶和修饰酶是两种不同的蛋白，它们各自独立行使自己的功能；而在另一些系统中，两种功能由同一种限制-修饰酶的不同亚基，或是同一亚基的不同结构域来执行。

传统上将限制性内切酶按照亚基组成、酶切位置、识别位点、辅助因子等因素划分为三大类。然而，蛋白测序的结果表明，限制性内切酶的变化多种多样，若从分子水平上分类，则应当远远不止这三种。

I 型限制性内切酶是一类兼有限制性内切酶和修饰酶活性的多个亚基的蛋白复合体。它

们在识别位点很远的地方任意切割 DNA 链。以前人们认为 I 型限制性内切酶很稀有，但现在通过对基因组测序的结果发现这一类酶其实很常见；尽管 I 型酶在生化研究中很有意义，但由于不产生确定的限制片段和明确的跑胶条带，因而不具备实用性。

II 型酶在其识别位点之中或临近的确定位点特异地切开 DNA 链。它们产生确定的限制片段和跑胶条带，因此是三类限制性内切酶中唯一用于 DNA 分析和克隆的一类。II 型限制性内切酶由一群性状和来源都不尽相同的蛋白组成，因而任意一种限制性内切酶的氨基酸序列可能与另一种限制性内切酶的氨基酸序列截然不同。实际上，从已知的情况上看，这些酶很可能是在进化过程中各自独立产生的，而非来源于同一个祖先。

II 型限制性内切酶中最普遍的是象 HhaI、HindIII 和 NotI 这样在识别序列中进行切割的酶。这一类酶是构成商业化酶的主要部分。大部分这类酶都以同二聚体的形式结合到 DNA 上，因而识别的是对称序列；但有极少的酶作为单聚体结合到 DNA 上，识别非对称序列。一些酶识别连续的序列（如 EcoRI 识别 GAATTC）；而另一些识别不连续的序列（如 Bgl I 识别 GCCNNNNNGGC）。限制性内切酶的切割后产生一个 3'羟基端和一个 5'磷酸基团。它们的活性要求镁离子，而相应的修饰酶则需要 S-甲硫氨酸腺苷的存在。这些酶一般都比较小，亚基一般都在 200-300 个氨基酸左右。

另一种比较常见的酶是所谓的 IIS 型酶，比如 FokI 和 AlwI，它们在识别位点之外切开 DNA。这些酶的大小居中，约为 400-650 个氨基酸左右；它们识别连续的非对称序列，有一个结合识别位点的域和一个专门切割 DNA 的功能域。一般认为这些酶主要以单体的形式结合到 DNA 上，但与临近的酶结合成二聚体，协同切开 DNA 链。因此一些 IIS 型的酶在切割有多个识别位点的 DNA 分子时，活性可能更高。

第三种 II 型限制性内切酶（有时也被称为 IV 型限制性内切酶）是一类较大的、集限制和修饰功能于一体的酶，通常由 850-1250 个碱基组成，在同一条多肽链上同时具有限制和修饰酶活性。有些酶识别连续序列，并在识别位点的一端切开 DNA 链；而另一些酶识别不连续的序列（如 Bcgl：CGANNNNNNTGC），并在识别位点的两端切开 DNA 链，产生一小段含识别序列的片段。这些酶的氨基酸序列各不相同，但其结构组成是一致的。他们在 N 端由一个负责切开 DNA 的功能域，这个域又和 DNA 修饰域连接；此外还有一到两个识别特异 DNA 序列的功能域构成 C 端，或以独立的亚基形式存在。当这些酶与底物结合时，它们或行使限制性内切酶的功能切开底物，或作为修饰酶将其甲基化。

III 型限制性内切酶也是兼有限制-修饰两种功能的酶。它们在识别位点之外切开 DNA 链，并且要求识别位点是反向重复序列；它们很少能产生完全切割的片段，因而不具备实用价值，也没有人将其商业化。

目前我们一般选用 II 型总有确定识别位点的酶作为分子生物学研究使用的酶，本酶切软件也基于这个前提而开发。

## 二、第二代测序的运用

高通量测序技术是对传统测序一次革命性的改变，一次对几十万到几百万条 DNA 分子进行序列测定，因此在有些文献中称其为下一代测序技术(next generation sequencing)足见其划时代的改变，同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序(deep sequencing)。

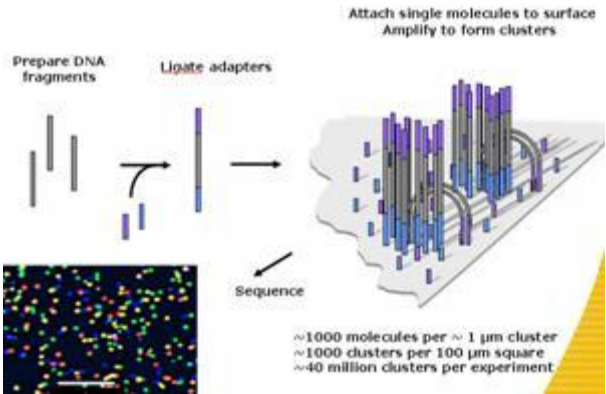
高通量测序技术是对传统测序一次革命性的改变，一次对几十万到几百万条 DNA 分子进行序列测定，因此在有些文献中称其为下一代测序技术(next generation sequencing)足见其划时代的改变，同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序(deep sequencing)。

自从 2005 年 454 Life Sciences 公司（2007 年该公司被 Roche 正式收购）推出了 454 FLX 焦磷酸测序平台（454 FLX pyrosequencing platform）以来，曾推出过 3730xl DNA 测序仪（3730xl DNA Analyzer）的 Applied BioSystem（ABI）这家一直占据着测序市场最大份额的公司的领先地位就开始动摇了，因为他们的拳头产品毛细管阵列电泳测序仪系列（series capillary array electrophoresis sequencing machines）遇到了两个强有力的竞争对手，一个就是罗氏公司（Roche）的 454 测序仪（Roche GS FLX sequencer），另一个就是 2006 年美国 Illumina 公司推出的 Solexa 基因组分析平台（Genome Analyzer platform），为此，2007 年 ABI 公司推出了自主研发的 SOLiD 测序仪（ABI SOLiD sequencer）。这三个测序平台即为目前高通量测序平台的代表。（见表一）。

| 公司名称                   | 技术原理                   | 技术开发者   | 商业模式                        |
|------------------------|------------------------|---|-----------------------------|
| Apply Biosystems (ABI) | 基于磁珠的大规模并行克隆连接 DNA 测序法 | 美国 Agencourt 私人基因组学公司 (APG)                   | 上市公司：销售设备和试剂获取利润            |
| Illumina               | 合成测序法                  | 英国 Solexa 公司首席科学家 David Bentley               | 上市公司：销售设备和试剂获取利润            |
| Roche                  | 大规模并行焦磷酸合成测序法          | 美国 454 Life Sciences 公司的创始人 Jonathan Rothberg | 上市公司：销售设备和试剂获取利润            |
| Helicos                | 大规模并行单分子合成测序法          | 美国斯坦福大学生物工程学家 Stephen Quake                   | 上市公司：2007 年 5 月首次公开募股 (IPO) |
| Complete Genomics      | DNA 纳米阵列与组合探针锚定连接测序法   | 美国 Complete Genomics 公司首席科学家 radoje drmanac   | 私人公司：投资额为 4650 万美元          |

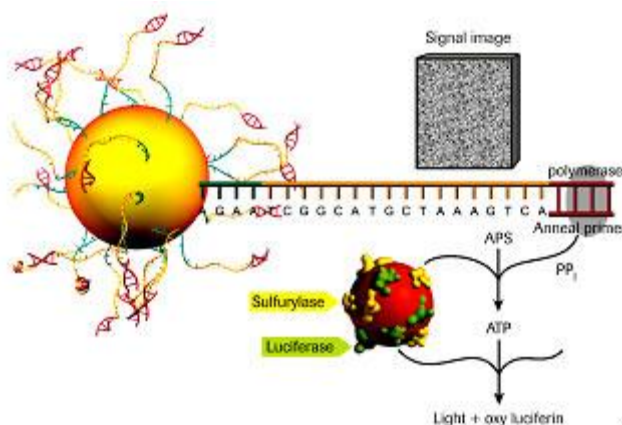
表一 主流测序平台一览

这些平台共同的特点是极高的测序通量，相对于传统测序的 96 道毛细管测序，高通量测序一次实验可以读取 40 万到 400 万条序列。读取长度根据平台不同从 25bp 到 450bp，不同的测序平台在一次实验中，可以读取 1G 到 14G 不等的碱基数，这样庞大的测序能力是传统测序仪所不能比拟的。尽管如此，在这项新的划时代的测序技术刚出现的时候，科学界对这项新技术却并不热衷。许多习惯用桑格技术的科学家怀疑新技术的准确度、阅读能力、成本消费、实用性。代理 Sanger 型测序硬件的经销商害怕其投资失败而首先提出了这些怀疑。



图一 在芯片上进行的测序：Illumina 测序平台

然而大多数人却忽略了一个事实，即桑格技术的普及最初也遇到同样的阻碍。桑格技术刚开发出来时，阅读能力很难超过 25bp，即使在 Fred Sanger 双脱氧终止法发明后也只达到 80bp，如今却达到了 750bp；而新发展的合成测序技术，应用焦磷酸测序方法，其阅读能力最初只有 100bp，推向市场 16 个月后增加至 250bp，随着技术的不断完善，目前已达到了 400bp，很快就接近桑格技术目前的水平。除了阅读能力外，能否以有限的成本用一台仪器产生足够数量的序列标记也是另一个需要改善的重要问题。这个问题已经被 Roche 公司解决了，应用他们的系统，仅花费阅读 35bp 或者更小片段的成本就能产生比 35bp 多 10 倍的序列标记。



图二 GS FLX 高通量测序方法原理示意图

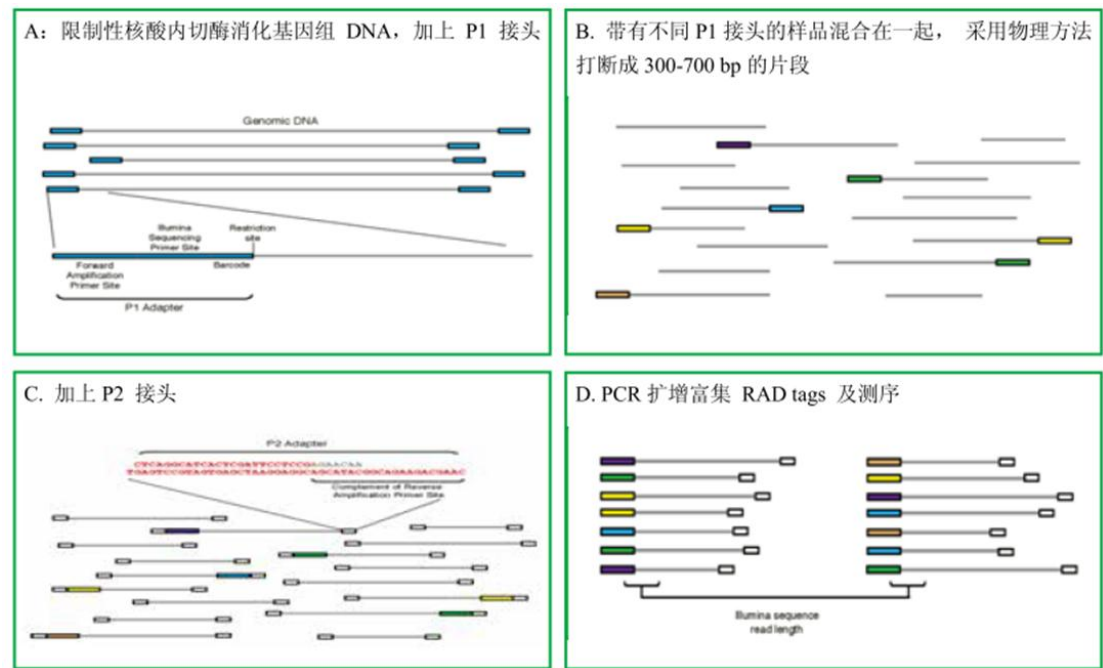
高通量测序可以帮助研究者跨过文库构建这一实验步骤,避免了亚克隆过程中引入的偏差。依靠后期强大的生物信息学分析能力,对照一个参比基因组(reference genome)高通量测序技术可以非常轻松完成基因组重测序(re-sequence),2007 年 van Orsouw 等人结合改进的 AFLP 技术和 454 测序技术对玉米基因组进行了重测序,该重测序实验发现的超过 75% 的 SNP 位点能够用 SNPWave 技术验证,提供了一条对复杂基因组特别是含有高度重复序列的植物基因组进行多态性分析的技术路线。2008 年 Hillier 对线虫 CB4858 品系进行 Solexa 重测序,寻找线虫基因组中的 SNP 位点和单位点的缺失或扩增。但是也应该看到,由于高通量测序读取长度的限制,使其在对未知基因组进行从头测序(novo sequencing)的应用受到限制,这部分工作仍然需要传统测序(读取长度达到 850 碱基)的协助。但是这并不影响高通量测序技术在全基因组 mRNA 表达谱, microRNA 表达谱, ChIP-chip 以及 DNA 甲基化等方面的应用。

### 三、酶切技术在第二代测序中的运用

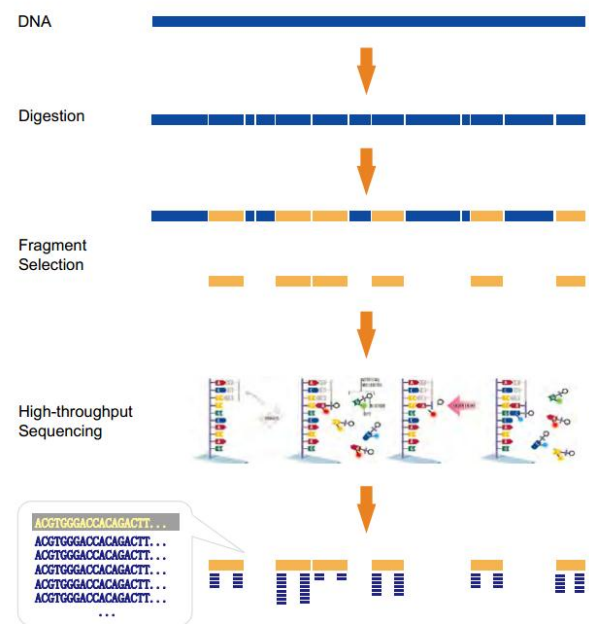
### 1、酶切技术在遗传图谱构建中的运用。

RAD (Restriction-site Associated DNA) 和 SLAF-seq (specific-locus amplified fragment sequencing) 都是与限制性核酸内切酶识别位点相关 DNA, 基于酶切的简化基因组测序, 对酶切获得的 tag 进行高通量测序, 大幅降低基因组的复杂度, 操作简便, 同时不受参考基因组的限制, 可快速鉴定出高密度的 SNP 位点。他们是基于 SNP 位点的分子标记技术, 性价比高、稳定性好, 可用于群体进化研究、遗传图谱构建、QTL 定位和辅助 scaffold 组装到染

染色体等领域。



图三 RAD-seq 文库构建流程（Baird 等，PLOS ONE，2008）

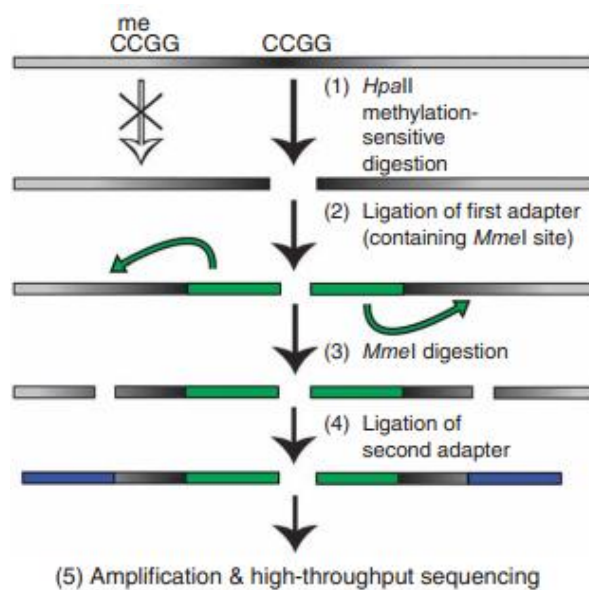


图四 SLAF-seq 文库构建流程（Xiaowen Sun 等，PLOS ONE，2013）

## 2、酶切技术在表观遗传学中的运用

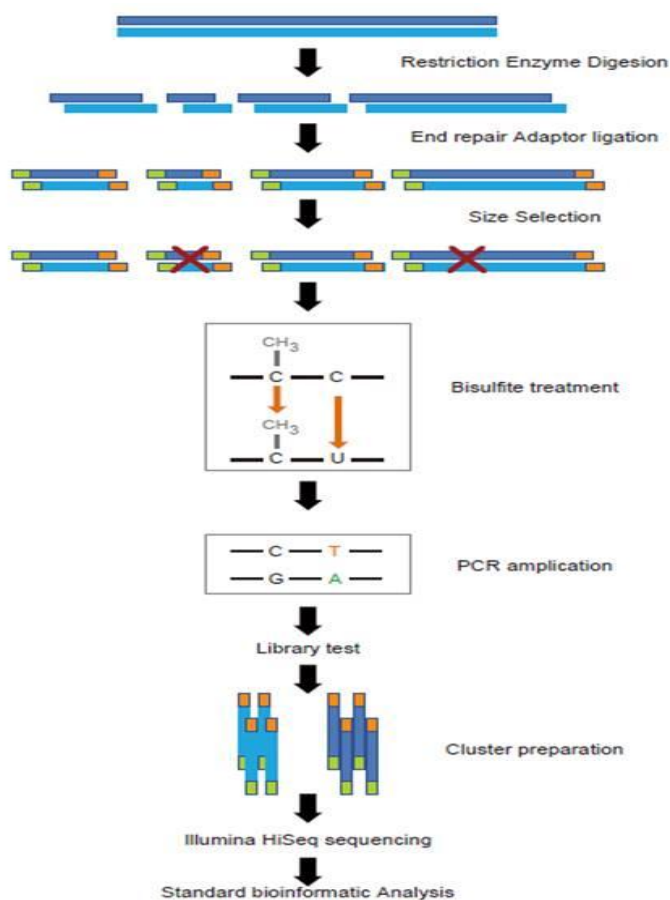
MRE-seq (Methylation-Sensitive Restriction Enzymes sequencing)利用甲基化高度敏感内切酶酶切基因组，然后将切碎的片段进行大规模测序，被测片段是无甲基化区域，而无序列覆盖的区域的 CG 就是甲基化位点。





图五 MRE-seq 实验原理示意图

RRBS(Reduced Representation Bisulfite Sequencing)利用限制性内切酶对基因组进行酶切，可以极大幅度的提高 CpG 区域的测序深度，与全基因组甲基化测序相比，测序量将大大减少，并在 CpG 岛、启动子区域和增强子元件区域达到更高精度的分辨率。同时，利用 RRBS 可以实现多个样本的比对基因组分析。

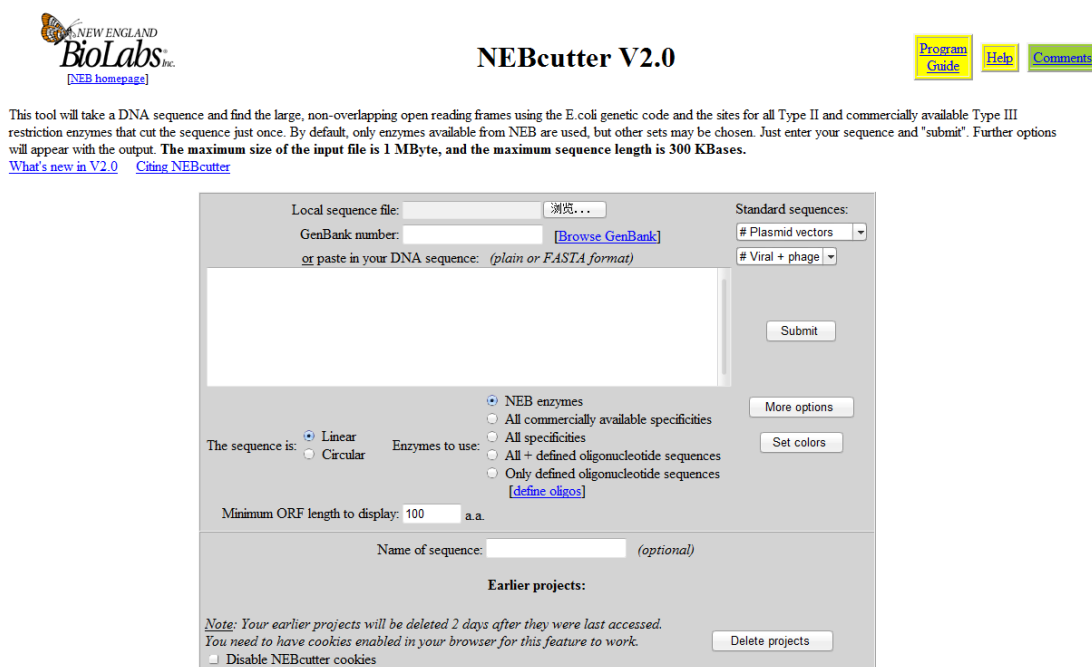


图六 RRBS-seq 流程图

## 研究内容

目前已经有一些相关的工具，大多数是知名的生物技术公司或者研究所开发的，有些是 web severer，有些是本地版的软件或者 R 包。比如 enzymex、restriction mapper、NEBCutter、Rebase、Mapper、Web Cutter 等等，一些被广泛运用的限制性酶切位点分析小工具可以查看网页：<http://www.helixnet.cn/uchome/space.php?uid=24114&do=blog&id=1616>，这些酶切分析工具做的都比较精巧，比如图这是网页版的酶切软件界面，一般可以输入单个小于 300k 的序列和小于 1M 的文件，能够识别单个酶或者多个酶联合时的酶切位点。

但是如果这些软件直接运用到我们的测序中来显然是不够的，不足之处在于：可以输入的序列有限，多个酶酶切时容易出现效率低，关键是不能保留黏端信息以及各个片段的序列，最需要改进的是没有合适的算法去帮助研究人员选择多大的片段范围比较适合自己领域的研究。



图七 NEB cutter 网页用户界面

而酶切技术在第二代测序中的运用非常广泛，包括遗传图谱的构建和 DNA 甲基化的检测。简化的遗传图谱构建是运用了酶切消化 DNA 实现片段化，对酶切获得的 tag 进行高通量测序，大幅降低基因组的复杂度，操作简便，同时不受参考基因组的限制，可快速鉴定出高密度的 SNP 位点。对于已经有参考基因组的物种来说，模拟酶切选择合适的片段大小尤为重要，原因有两个：（1）因为不同仪器的测序读长是不一样的，因此在建库过程中所需要的插入片段长度也有一定的限制，比如对于 Illumina 平台来说，DNA 插入片段的长度一般在 400-500bp 左右，如需测通的话，这个片段长度要进一步剪短至读长范围内；如果是 454 测序的话，一般插入片段长度可以达到 1Kb 以上的。（2）DNA 扩增需要合适的碱基含量(即 GC 含量合适)和碱基平衡性。因此，实验之前预测完全模拟酶切的效率、大小、GC 含量和碱基平衡性等对确保后续实验成功尤为重要，内切酶的筛选和模拟酶切的效果评估是优化实验方案的必要准备工作。

DNA 甲基化能关闭某些基因的活性，去甲基化则诱导了基因的重新活化和表达。DNA

甲基化能引起染色质结构、DNA 构象、DNA 稳定性及 DNA 与蛋白质相互作用方式的改变，从而控制基因表达。研究证实，CpG 二核苷酸中胞嘧啶的甲基化导致了人体 1/3 以上由于碱基转换而引起的遗传病。DNA 甲基化主要形成 5-甲基胞嘧啶 (5-mC) 和少量的 N6-甲基腺嘌呤 (N6-mA) 及 7-甲基鸟嘌呤 (7-mG)。在真核生物中，5-甲基胞嘧啶主要出现在 CpG 序列、CpXpG、CCA/TGG 和 GATC 中。由此可见，DNA 修饰是发生在特定碱基并且具有特定序列特征，而检测 DNA 甲基化测序技术 BS-seq 是对全基因组所有碱基测序，这样导致很大一部分碱基是无用的或多余的。目前运用酶切技术的 RRBS-seq，一定程度上克服了这种“浪费”，对人和小鼠采用 MspI 酶切然后选取 40-220bp 大小片段区域测序，只需要测基因组的 2% 左右的区域，却可以检测到全基因组 10% 的 CG 甲基化情况，并且大部分被检测的 CG 都是人们比较关注的启动子和 CpG 岛区域。同样，对于其他目前研究甚少的碱基修饰，将来运用相同的测序策略也许能实现相似的“物美价廉”效果。

综上所述，为融合酶切技术在第二代技术的充分和合理运用，开发出一个灵活的、全面的、高效的、友好的酶切模拟和评估软件是迫在眉睫。因此，我们有意开发一款名为“魔法酶切” (magic enzyme cutter) 的软件，旨在完成和实现以下目的：

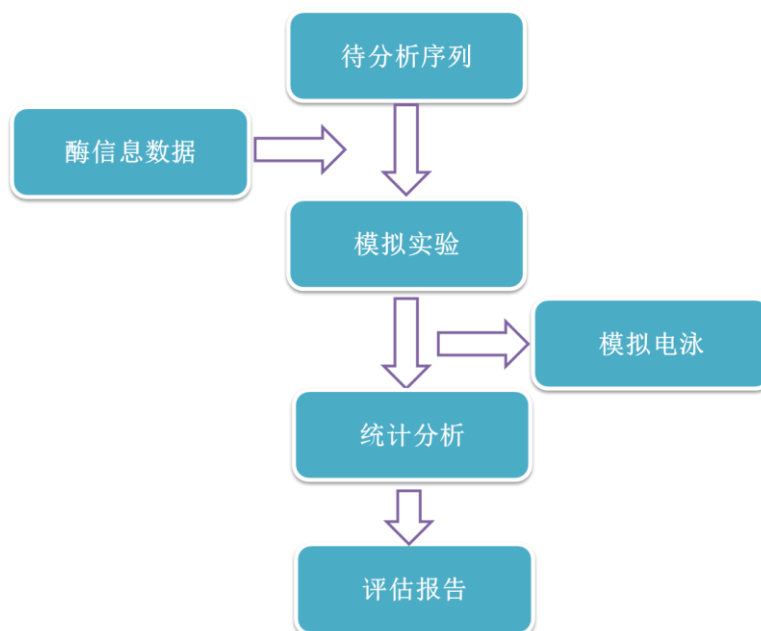
- (1) 大规模高效率的模拟酶切，得到每一个酶切片段的黏端信息和碱基序列，解决以往酶切工具的基因组大小限制问题。
- (2) 统计和显示酶切片段大小、GC 含量和碱基平衡性，以评估特定长度大小片段的扩增效率。
- (3) 以电泳图的形式展示模拟酶切结果，形象生动的展示酶切效果。
- (4) 计算和统计特定长度范围上的特定序列的比例，为特定序列靶向测序特供依据和支持。
- (5) 使用最优解算法优化酶切的区域范围的选择，使所选择的区域范围对目标碱基或特征序列测序效能最高。

最终将上面的结果以网页的形式汇总成评估报告，方便研究者和使用者一目了然地知道模拟酶切效果。

以上工作将通过 perl 语言、R 语言和 HTML 语言编程完成。

## 研究方法

该软件的结果如下图所示：



图八 软件结构图

当选择了目标酶和基因组，并设定合适的参数之后，即可以开始运行模拟酶切。模拟酶切之后模拟电泳和统计分析，最终生成网页 HTML 版的评估报告。

### 1、字符串匹配算法

模拟酶切软件算法的实质就是通过限制性内切酶的识别序列对消化的序列进行匹配的过程，我们软件中要考虑到用什么样的字符匹配算法，大家知道的比较著名的匹配算法是 Hash 算法，KMP 算法和 BM 算法（Boyer-Moore 算法），下面我简要介绍一下：

Hash 算法的核心思想是样本总空间思想，对于长度为  $n$  的字符串，字符的 ASCII 码值为  $1 \sim 255$ ，则样本总空间是  $255^n$ ，利用样本总空间是绝对没有冲突的可能，但是  $n$  大于一定的数值就没有办法实现，于是采用 Hash 函数算法： $\text{Hash}(X_i) = X_i \cdot B_i^{R_i}$ ， $B_i$  可以是素数序列也可能为固定为某个值， $R_i$  可以是自然数序列也可以固定为某个值。这些系数的选择要根据不同问题而定。尽管如此，字符串的长度还是有一定的限制。实现算法需要  $(m-n)$  步。KMP 算法的基本思想是利用已经匹配的结果来简化算法的复杂度。当字符串中的第  $i$  个字符出现不匹配时，字符串就回到  $\text{next}(i)$  个位置，主字符串的比较窗口和比较位置都发生了相应的变化。如果子字符串为  $S[i]:\text{asasaadtas}$

$\text{next}[i]:1112342112$

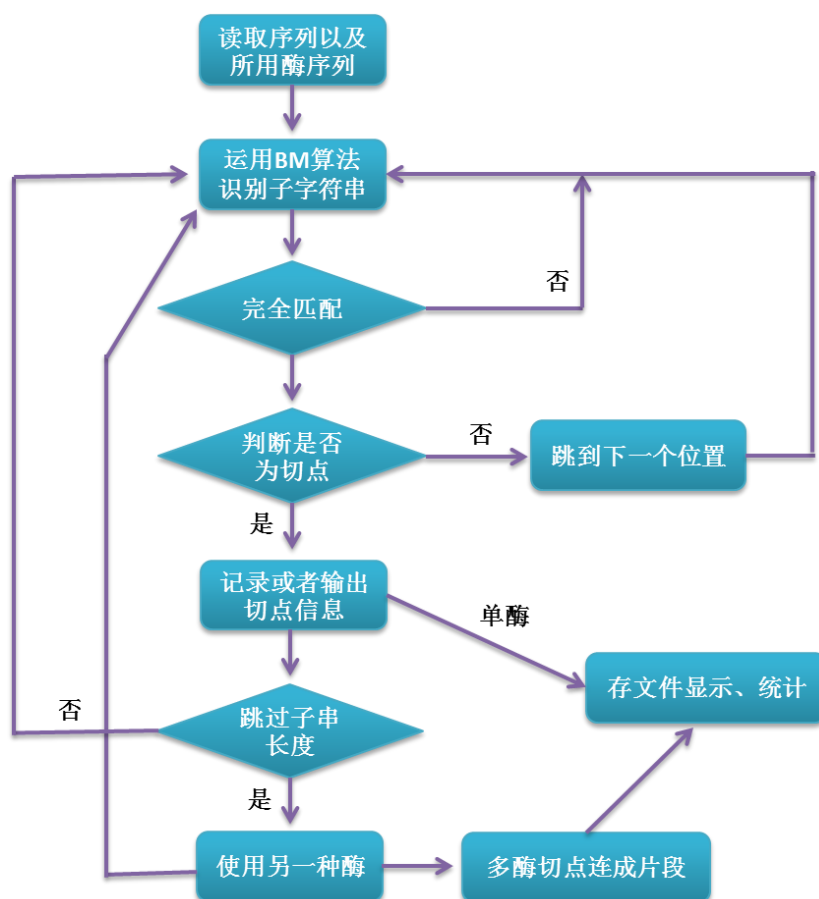
主字符串  $M[i]:\text{asasasetasaaadtastasasasaa...}$

我们的假设匹配的窗口标志是  $p$ ，即匹配的内容是  $M[p] \dots M[n]$  为子字符串长度， $M[i]$  为主字符串组。当匹配到主字符串  $M[6]$  为  $s$ ，即  $S[6]$  为  $a$ ，由于不匹配到下一次匹配时，主字符串的窗口标志从  $p=1$  跳到  $p=3=p+i-\text{next}[6]$ ：即主字符串的第二个  $a$ ，子字符串退到第  $\text{next}[6]$  个位

置：即子字符串中的第 2 个 s。字符串中比较位是  $6=p+next[6]-1$ ，也就是用  $M[6]$ 跟  $S[4]$ 比较。该算法的运算复杂度低，但是匹配步数远大于  $m-n$ 。

BM 算法的思想和 KMP 思想类似，但是方向相反。它是利用字符串还没有匹配的内容来简化算法的复杂度。如果没有匹配成功的字符（ $M[i]$ ，对应于子字符串  $S[j]$ ）不在子字符串还没有匹配过的内容中，主字符串的窗口后退到主字符串中该字符与子字符串中该位置对齐。可以看出该算法的运算复杂度同样大于  $m-n$ ，但是计算复杂度明显降低了。

我们使用的算法是 Boyer-Moore 算法即 BM 算法，BM 算法是一种非常高效的精确字符串匹配算法（只是一个启发式的字符串搜索算法）。它由 Bob Boyer 和 J Strother Moore 设计于 1977 年。此算法仅对搜索目标字符串（关键字）进行预处理，而非被搜索的字符串。虽然 Boyer-Moore 算法的执行时间同样线性依赖于被搜索字符串的大小，但是通常仅为其它算法的一小部分：它不需要对被搜索的字符串中的字符进行逐一比较，而会跳过其中某些部分。通常搜索关键字越长，算法速度越快。它的效率来自于这样的事实：对于每一次失败的匹配尝试，算法都能够使用这些信息来排除尽可能多的无法匹配的位置。我们开发的软件“魔法酶切”的算法结构如下：



图九 酶切软件基本算法

## 2、智能优化算法。

粒子群优化(PSO) 是一种新兴的基于群体智能的启发式全局搜索算法，具有易理解、易实现、全局搜索能力强等特点，倍受科学与工程领域的广泛关注，成为发展最快的智能优化算法之一。PSO 算法最初是为了图形化的模拟鸟群优美而不可预测的运动。而通过对动物社

会行为的观察，发现在群体中对信息的社会共享提供一个演化的优势，并以此作为开发算法的基础。通过加入近邻的速度匹配、并考虑了多维搜索和根据距离的加速，形成了 PSO 的最初版本。之后引入了惯性权重  $w$  来更好的控制开发(exploitation)和探索(exploration)，形成了标准版本。

PSO 算法是基于群体的，根据对环境的适应度将群体中的个体移动到好的区域。然而它不对个体使用演化算子，而是将每个个体看作是  $D$  维搜索空间中的一个没有体积的微粒(点)，在搜索空间中以一定的速度飞行，这个速度根据它本身的飞行经验和同伴的飞行经验来动态调整。第  $i$  个微粒表示为  $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ ，它经历过的最好位置（有最好的适应值）记为  $P_i = (p_{i1}, p_{i2}, \dots, p_{iD})$ ，也称为  $pbest$ 。在群体所有微粒经历过的最好位置的索引号用符号  $g$  表示，即  $P_g$ ，也称为  $gbest$ 。微粒  $i$  的速度用  $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$  表示。对每一代，它的第  $d$  维( $1 \leq d \leq D$ )根据如下方程进行变化：

$$vid = w \cdot vid + c1 \cdot rand() \cdot (pid - xid) + c2 \cdot Rand() \cdot (pgd - xid) \quad (1a)$$

$$xid = xid + vid \quad (1b)$$

其中  $w$  为惯性权重(inertia weight)， $c1$  和  $c2$  为加速常数(acceleration constants)， $rand()$  和  $Rand()$  为两个在  $[0,1]$  范围里变化的随机值。

此外，微粒的速度  $V_i$  被一个最大速度  $V_{max}$  所限制。如果当前对微粒的加速导致它的在某维的速度  $vid$  超过该维的最大速度  $v_{max,d}$ ，则该维的速度被限制为该维最大速度  $v_{max,d}$ 。

对公式(1a)，第一部分为微粒先前行为的惯性，第二部分为“认知(cognition)”部分，表示微粒本身的思考；第三部分为“社会(social)”部分，表示微粒间的信息共享与相互合作。

“认知”部分可以由 Thorndike 的效应法则(law of effect)所解释，即一个得到加强的随机行为在将来更有可能出现。这里的行为即“认知”，并假设获得正确的知识是得到加强的，这样的模型假定微粒被激励着去减小误差。

“社会”部分可以由 Bandura 的替代强化(vicarious reinforcement)所解释。根据该理论的预期，当观察者观察到一个模型在加强某一行为时，将增加它实行该行为的几率。即微粒本身的认知将被其它微粒所模仿。

PSO 算法使用如下心理学假设：在寻求一致的认知过程中，个体往往记住自身的信念，并同时考虑同事们的信念。当其察觉同事的信念较好的时候，将进行适应性调整。

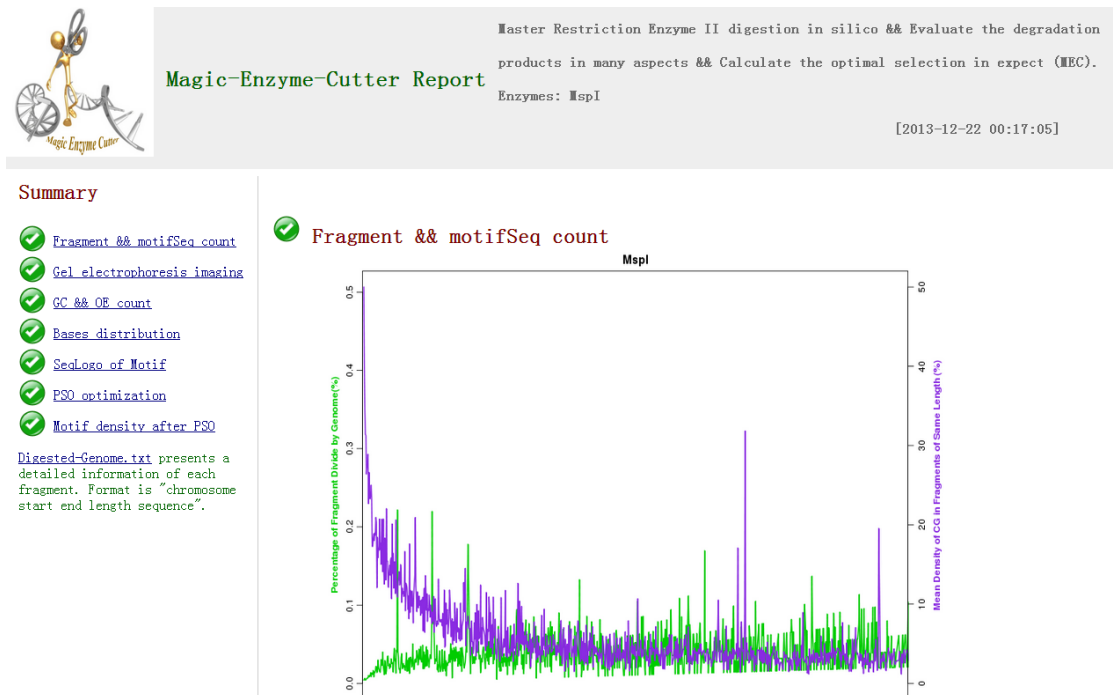
标准 PSO 的算法流程如下：

1. 初始化一群微粒（群体规模为  $m$ ），包括随机的位置和速度；
2. 评价每个微粒的适应度；
3. 对每个微粒，将它的适应值和它经历过的最好位置  $pbest$  的作比较，如果较好，则将其作为当前的最好位置  $pbest$ ；
4. 对每个微粒，将它的适应值和全局所经历最好位置  $gbest$  的作比较，如果较好，则重新设置  $gbest$  的索引号；
5. 根据方程(1)变化微粒的速度和位置；
6. 如未达到结束条件（通常为足够好的适应值或达到一个预设最大代数  $G_{max}$ ），回到 2)。

我们需要用这个算法快速计算出目标碱基或者特征序列的最佳酶切片段范围，从而实现测量最少的基因组范围但得到最多的目标碱基或特征序列，并且这些序列或碱基与相关生命科学研究（如 DNA 甲基化）密切相关，最终达到经济适用的状态。

# 研究结果

软件运行成功之后得到的结果以网页形式展现，测试数据的结果如下：



测序运行的参数为：

```
Magic-Enzyme-Cutter -i ref.fa.gz -o $pwd_dir/test-result -R $R
```

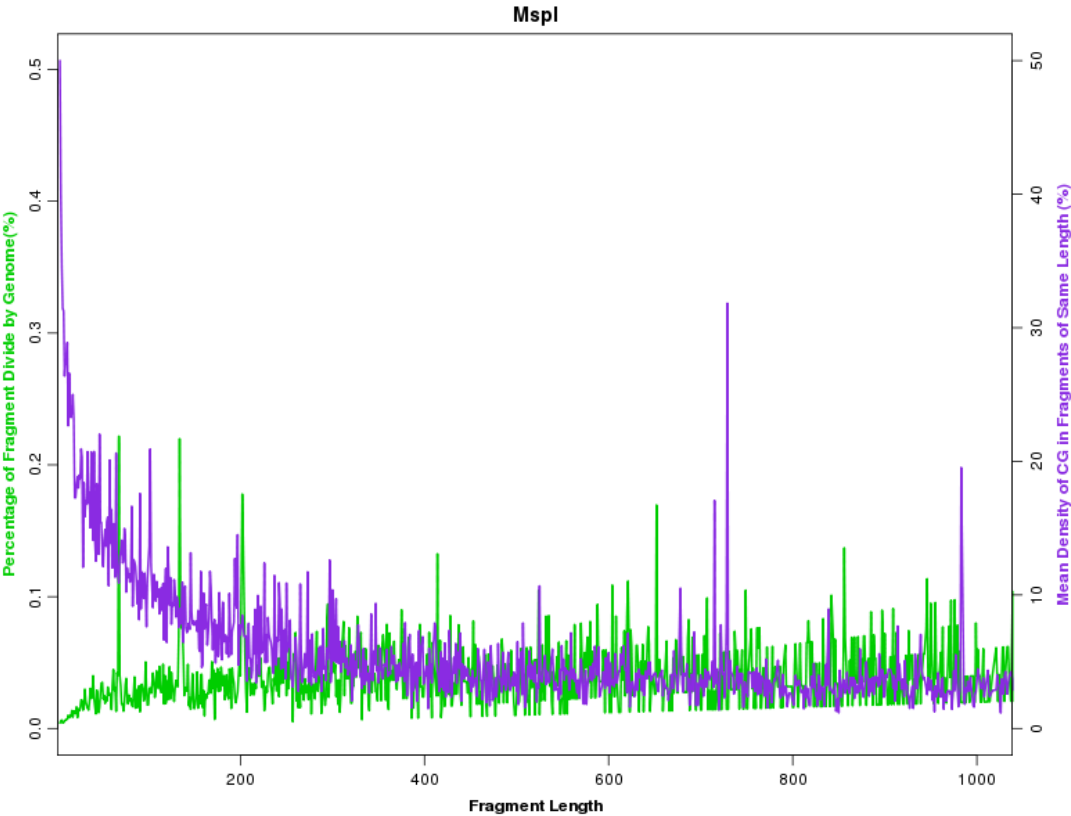
| 参数 | 参数值（默认值）           | 解释                       |
|----|--------------------|--------------------------|
| -i | ref.fa.gz          | Fasta 格式的参考基因组           |
| -o | ./test/test-result | 输出结果文件夹                  |
| -m | CG                 | 感兴趣的 motif               |
| -R | Path/R             | R 路径                     |
| -s | 40                 | 选取酶切片段最小长度               |
| -e | 1000               | 选取酶切片段最大长度               |
| -n | 200                | PSO 最优化的 motif 密度的最小片段范围 |
| -x | 500                | PSO 最优化的 motif 密度的最大片段范围 |
| -f | 4                  | Motif 上下游碱基数目            |
| -g | 100                | 碱基分布密度值                  |
| -c | F                  | 是否改变图像视觉效果（F 表示否，T 表示是）  |
| -t | 50                 | 凝胶电泳图透明度                 |

运行时间为：

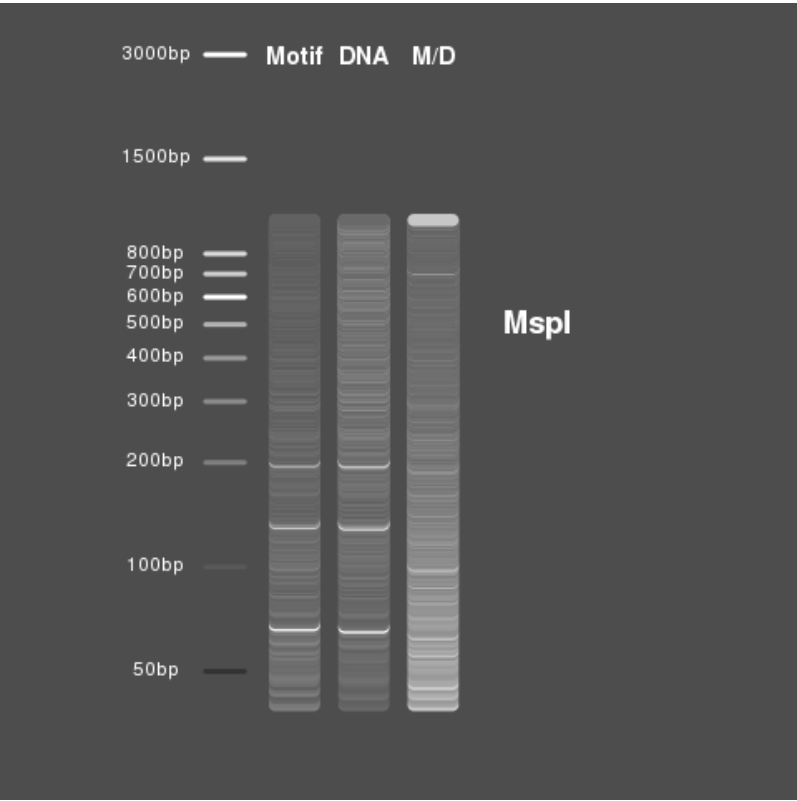
```
real    0m13.394s
user    0m10.265s
sys     0m0.898s
```

图形展示：

1、模拟酶切片段及 motif 分布图

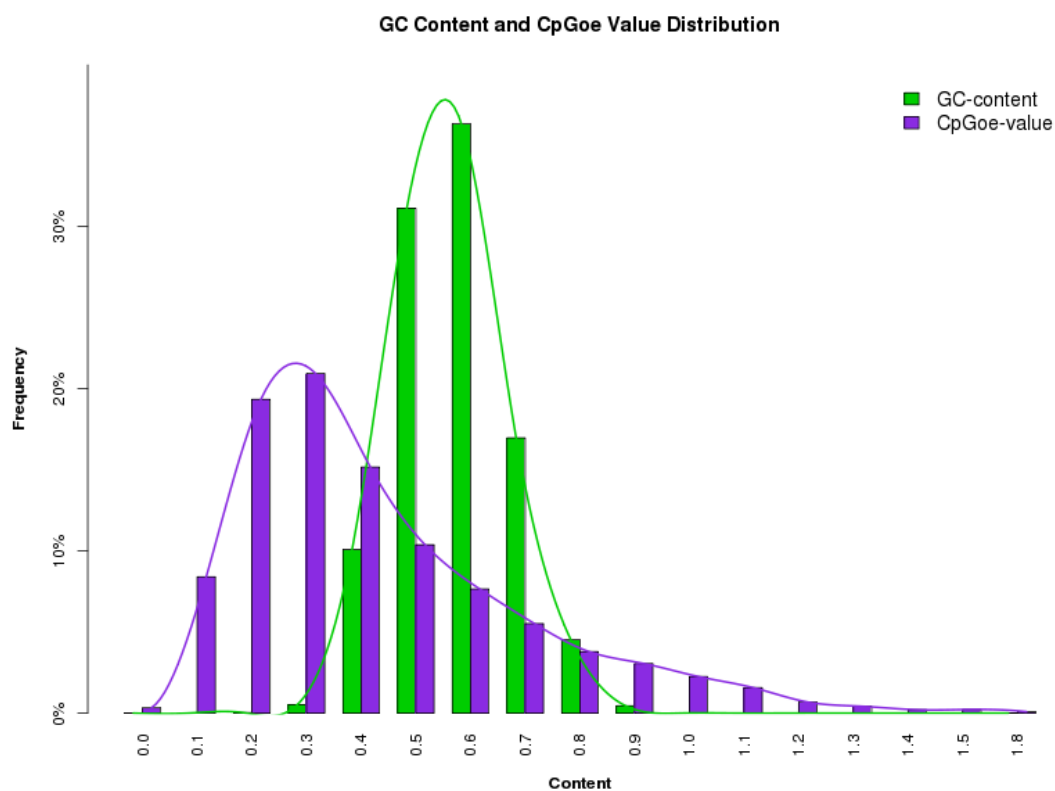


2、模拟酶切胶图。

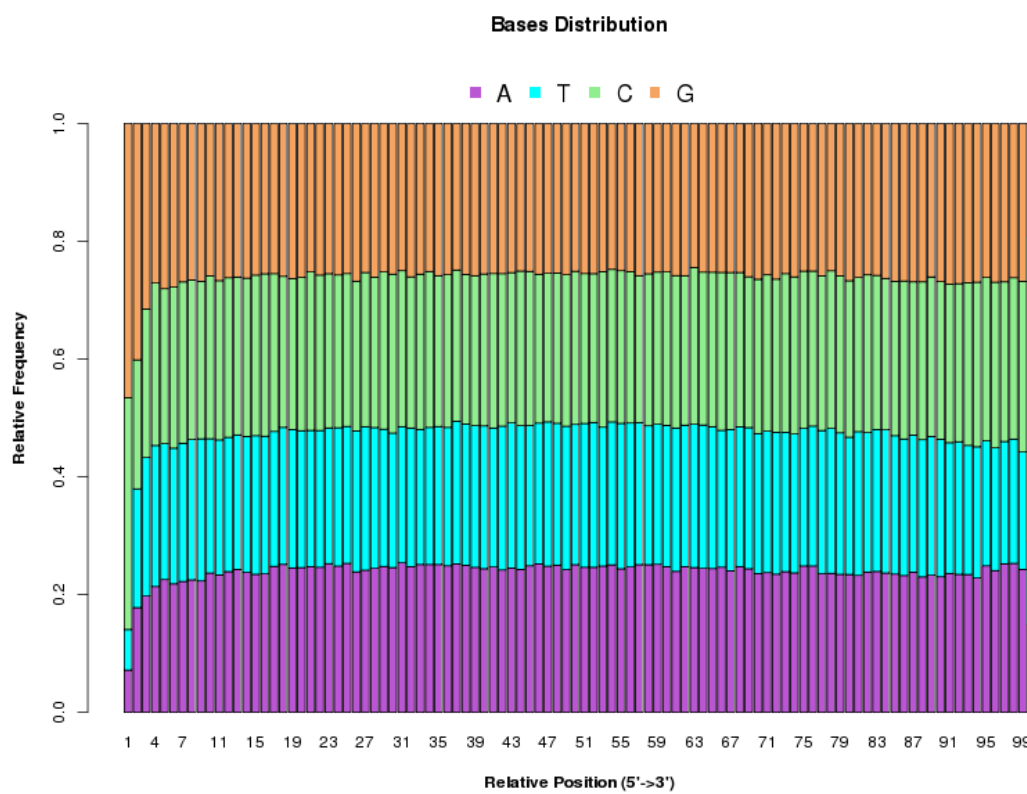




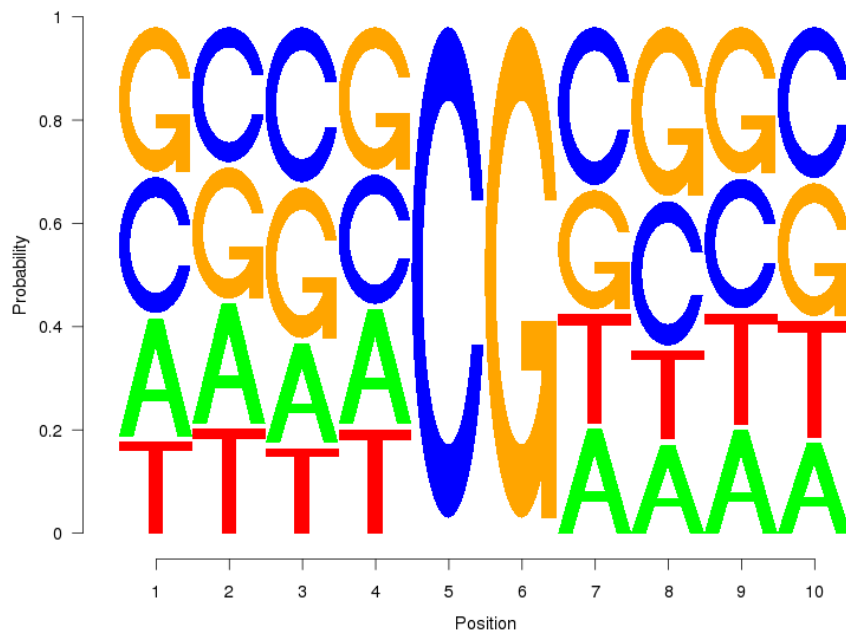
### 3、GC & OE 分布图:



### 4、碱基分布图:



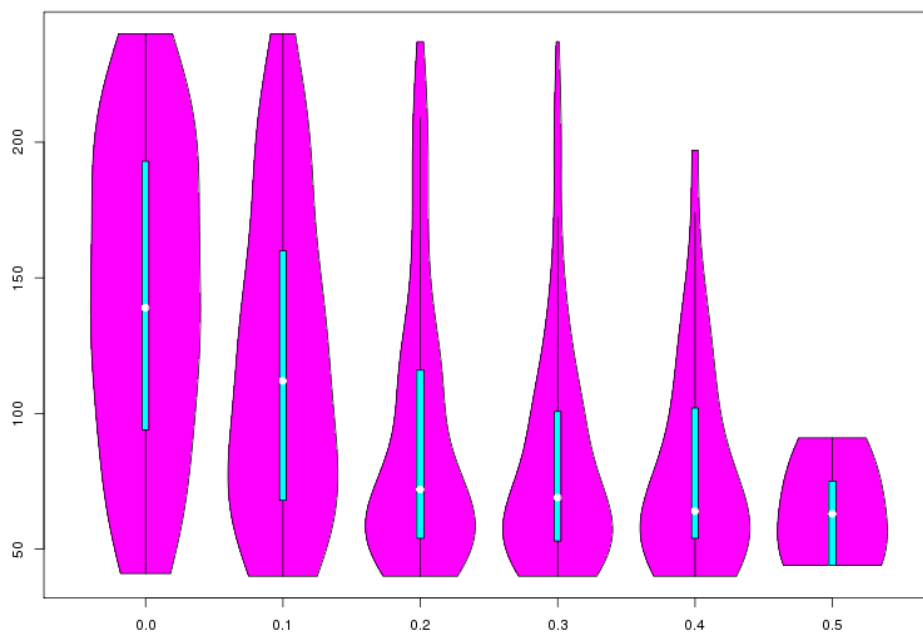
### 5、目标 motif 及两侧碱基的 seqLogo 图



### 6、PSO 最优化得到的起始长度

| Start | End   |
|-------|-------|
| 40bp  | 240bp |

### 7、PSO 最优化得到的 motif 密度分布图



## 8、酶切文件显示

Digested-Genome.txt presents a detailed information of each fragment. Format is "chromosome start end length sequence".

格式：染色体 起始位置 终止位置 片段长度 碱基序列

chrR 1 212 212

```
GATCTGATAAGTCCCAGGACTTCAGAAGAGCTGTGAGACCTTGGCCAAGTCACTTCCTCCTTCAGGAA
CATTGCAGTGGGCCTAAGTGCCTCCTCTCGGGACTGGTATGGGGACGGTCATGCAATCTGGACAACATTCAC
CTTTAAAAGTTTATTGATCTTTTGTGACATGCACGTGGGTTCACAGTAGCAAGAACTAAAGGGTCGCAGGC
```

chrR 213 1357 1145

```
CGGTTTCTGCTAATTTCTTAATTCCAAGACAGTCTCAAATATTTTCTTATTAATTCCTGGAGGGAGGC
TTATCATTCTCTCTTTTGGATGATTCTAAGTACCAGCTAAAATACAGCTATCATTCTTTTCTTGATTGGGAG
CCTAATTTCTTTAATTTAGTATGCAAGAAAACCAATTTGGAAATATCAACTGTTTTGGAAACCTTAGACCTAGG
TCATCCTTAGTAAGATCTTCCATTTATATAAATACTTGCAAGTAGTAGTGCCATAATTACCAAACATAAAGCCA
ACTGAGATGCCCAAAGGGGGCCACTCTCCTTGCTTTTCTCCTTTTTAGAGGATTTATTTCCATTTTTCTTAA
AAAGGAAGAACAACTGTGCCCTAGGGTTTACTGTGTCAGAACAGAGTGTGCCGATTGTGGTCAGGACTCC
ATAGCATTTCACCATGAGTTATTTCCGCCCTTACGTGTCTCTTTCAGCGGTCTATTATCTCCAAGAGGGC
ATAAACTGAGTAAACAGCTCTTTATATGTGTTTCTGGATGAGCCTTCTTTAATTAATTTGTTAAGGG
ATTTCTCTAGGGCCACTGCACGTCATGGGGAGTCACCCCAGACACTCCCAATTGGCCCCTTGTCACCCAG
GGGCACATTTAGCTATTTGTAAACCTGAAATCACTAGAAAAGGAATGTCTAGTGACTTGTGGGGGCAAG
GCCCTTGTTATGGGGATGAAGGCTCTTAGGTGGTAGCCCTCCAAGAGAATAGATGGTGAATGTCTCTTTCA
GACATTAAAGGTGTCAGACTCTCAGTTAATCTCTCCTAGATCCAGGAAAGGCCTAGAAAAGGAAGGCCTGA
CTGCATTAATGGAGATTCTCTCCATGTGCAAAATTTCTCCACAAAAGAAATCCTTGCAGGGCCATTTAATG
TGTTGGCCCTGTGACAGCCATTTCAAATATGTCAAAAAATATATTTTGGAGTAAAATACTTTCAATTTCTTTC
AGAGTCTGCTGTCGTATGATGCCATACCAGAGTCAGGTTGGAAAGTAAGCCACATTATACAGCGTTAACCTA
AAAAAACAAAAAATGTCTAACAAGATTTTATGGTTTATAGAGCATGATTCCC
```

