学校代码：10246
学号：032019029

硕 士 学 位 论 文

细菌基因组内分类特异性寡核苷酸重复序列的研究

院　　系　理论生命科学研究中心
复旦大学物理系

专　　业　计算生物学
凝聚态物理

姓　　名　夏 立

指 导 教 师　郝柏林教授

完 成 日 期　2006年4月10日

指导小组成员名单

郝柏林	教授
卢宝荣	教授
方海平	教授
张文驹	教授

目　　录

摘要

用计算生物学方法，我们对已测序的细菌完全基因组中的寡核苷酸重复序列作了详细研究。虽然重复序列在细菌基因组中所占比重不大，但它们仍然普遍存在，并且与细菌的生物许多功能密切相关。我们的研究表明，有些寡核苷酸重复序列在保持较高复制份数的同时具有很强的分类特异性。普遍在12个碱基的长度左右，一些特殊的寡核苷酸序列就已经可以成为细菌属一级的特征序列。更长的寡核苷酸序列会更特异，往往能够成为细菌种，甚至是菌株的标签序列。我们在古细菌和真细菌中，分别给出了一些不同分类特异级别的单例。由于计算结果数据量庞大，很不直观，为了更好的共享研究结果，我们搭建了TSOR服务器，提供即时的对于特异性寡核苷酸重复序列的查询，论文中给出了查询的范例。同时，这样的TSOR序列由于同时具有在基因组内较高的复制份数和限制分布在某一分类的细菌中的双重特征，使他们成为设计具有偏向性的全基因组扩增引物的很好候选。

关键词 细菌基因组重复序列数据服务器

ABSTRACT

Using the computational approach, we studied the oligonucleotides repeats in current available bacterial whole genomes. Though, repeats only count for a small portion in bacterial genomes, they still prevail. Our study shows, some of these oligonucleotides have a large copy number in genomes while maintain its taxon specificity. Generally, a length larger than 12 is enough to make a oligonucleotides repeats genus-specific. Longer oligonucleotides will become more specific and be the species or strain marker sequences. We show here some examples in archaea and bacteria with different specific taxon levels. As we have a large volume of computational results, we make it available online by our TSOR server.It deals with user's query and in this thesis we give examples on how to use this server. Moreover as these TSOR sequences are both specific and highly repeated, they would become possible nice candidate for biased microbial community genomes amplification.

Keywords Bacteria Genomes Repeats Database

第一章　绪论

1.1 细菌基因组进展

细菌生物（真细菌和古细菌）是生物圈的重要组成部分。自从三十多亿年前，细菌作为最古老的生物诞生于地球，到现在，它仍是地球上分布最广，适应性最强的生物[1]。作为个体的细菌是相对弱小的生物，但他们的集体作用却巨大的影响着地球生境和人类生活。有些细菌种类因为能引发严重的人类疾病而被我们所熟知。这其中包括导致某种流感的*Haemophilus influenzae*，导致肺炎的*Streptococcus pneumoniae*，导致某些事物中毒的*Staphylococci*属等。但有更多的细菌是对人类友善和有利的。其中就有很多作为分解者的细菌，生存在土壤，江河湖海底，把生物体中的化学元素以无机物的形式归还到自然环境中去。如果没有这些分解者细菌，所有真核生物必将灭绝，而相反细菌仍可以自由的生活下去。

正是因为细菌对于人类健康，生态环境和能源的重要性，并且细菌的基因组相对于真核生物而言相对较小的原因，早在1996年开始，就开始对陆续对细菌基因组进行测序，以便于对其进行更深入的研究。*Escherichia coli* K-12是最早开始测序的细菌全基因组[2]，而*Haemophilus influenzae* Rd是最早完成的细菌全基因组[3]。此后，随着全基因组散弹枪测序(Whole Genome Shotgun Sequencing，WGS)的引入和资金投入的增加，细菌基因组的测序速度和规模每年都在增加。截止2006年4月7日，共有被测序的细菌全基因组319个，包括26个古细菌(Archaea)和293个真细菌(Eubacteria)[4]。其中，对于人类肠道致病细菌最集中的*Proteobacteria*门中，已有多达64个菌株被测序。

与测序工作的进展相一致，基于细菌基因组的研究工作也取得了相当的进展。首先，产生了很多面向细菌基因组的找基因程序，包括Glimmer[5],GeneMark[6]等。由于细菌基因组相对于真核生物基因组而言结构简单，这些程序已具有非常高的准确率，一般对基因的识别率可达90%以上。其次，基于WGS方法的细菌基因组测序也发展迅速，随着拼接程序的完善，现在利用WGS方法，可以在一些大规模测序中心一天完成一个细菌的测序工作。最后，对于基因组的比较基因组研究(Comparatitive Genomics)，为分子生物学，遗传学提供大量的基础知识，同时利用基因组序列中的微卫

星重复(Small Satellite Repeats, SSR),串行重复(Tandem Repeats, TR)或其他一些特征序列作为遗传试验中的Marker已经成为遗传学研究中的重要工具[7]，对于DUS(DNA uptake signal)[8],χ-site[2]等特征序列的研究更是帮助揭示了很多重组过程中的分子机制。因此基因组学可谓是介与分子功能和遗传信息之间的密码学，随着更多的生物基因组将被测序，它的重要性不言喻。

1.2 细菌基因组中重复序列

在对于细菌基因组的研究中，对于细菌基因组中的重复序列的研究是其重要的一部分。基因组重复序列可以定义为与基因组中的其他序列共有很大相似性的序列。细菌基因组中的重复序列可以大致分为两种类型，低复杂度串行重复序列(Lowcomplexity TR)和较长的重复序列。低复杂度串行重复序列常指重复单元长度为1个核苷酸到5个核苷酸左右，头尾相接排列的多次出现的重复序列。较长的重复序列则包括转座元素(Transposable Elements)，迷你卫星序列(Minisatellite)，大规模串行重复序列(TR) 和间隔分布的重复序列(Spaced Repeats)。

虽然细菌基因组中的重复序列类型在真核生物基因组中都能找到对应，但是各种重复序列在基因组中所占比例及数量是不同的。

- 第一，相对于真核生物基因组而言，细菌基因组比较简洁，90%以上为编码区[9]，重复序列在细菌基因组中一般只占到1-2%。而在真核生物中，非编码区的比例要大得多，酵母中为28%，而在人类中为97-98%[10]，因为非编码区多为重复序列，所以真核生物的重复序列要多得多。

- 第二，简单重复序列广泛分布于真核生物基因组中，占所有重复序列总量中的绝大多数，如人类基因组中的Alu序列等[11]，可以到达几百kb。相反地，在细菌基因组中，简单重复序列所占比重很小，几乎很少出现。相对而言较长的重复序列的比重要比真核生物高很多。然而我们必须小心这样的比较，因为真核生物和细菌基因组的大小完全不同，其中基因大小的重复可能与细菌基因组中较长重复相当。然而基因大小的重复往往在真核生物中已不再被认为仅仅是重复序列。

- 第三，真核基因组的较长序列重复往往是大规模的转座子，如植物基因组中的Mite，Sine，Line序列等[12]。而在细菌基因组中，虽然转座子和插入序列(Insertion Sequence, IS)的分布也很广泛，但其总量不大，较长重复序列的主要成分是间隔分布的重复序列和大规模的串行重复序列。

当然除了以上三点不同外，真核生物和细菌基因组在重复序列上还有很多其他不同之处。

重复序列还可以按重复单元的匹配程度分为严格重复与不严格重复。每个重复单元的每次出现都严格一致就是严格重复，每次出现可以不一致就是不严格重复序列，或者叫做渐进重复序列(Approximated Repeats)。因为这些重复序列往往带有生物功能，所以很多用来寻找这些重复序列的软件也应运而生。常见的寻找不同重复序列的软件有STAR，STRING，TRF，SRF，Adplot，Reputer等[13–18]。

1.3 寡核苷酸重复序列

虽然细菌基因组中重复序列类型五花八门，在这篇论文中我们的主要研究对象是严格重复的寡核苷酸序列，重复单元长度从9个核苷酸到24个核苷酸。这样的重复序列在细菌基因组中的分布很广泛，并且往往蕴含着生物功能意义。早在1996年，第一个细菌全基因组*Haemophilus influenzae*测序完成时，就有工作对其中的频繁出现的寡核苷酸，也就是我们所谓的寡核苷酸重复序列进行了研究[19]。发现表明这些频繁出现的寡核苷酸序列分成三种类型。一是，Uptake Signal Sequences(USS)序列。这种USS序列也同样在*Pasteurellaceae*属的其他菌株中被发现[20]。二是，Multiple Tetranucleotide Interations，在编码区前段，重复次数的改变从而改变被表达的蛋白。三是，Intergenic Dyad Sequences(IDS)，在非编码区，可能与DNA的复杂二级结构有关。此后随着细菌全基因组数据的增加，同样的研究被推广到更多的基因组上，并与比较基因组学的方法相结合。文献中，人们对*Streptococci*，*Bacilli*等属基因组中的寡核苷酸重复序列做了研究，结果也表明他们与细菌的生物功能密切相关[21, 22]。

总而言之，在细菌基因组中，其中的高频度的寡核苷酸重复序列大概会有以下生物功能：

- 第一，作为信号序列(Biological Signals)，在生物功能中起到重要作用，譬如Ecoli基因组中的χ-site[2]，作为内切酶的识别位点；又譬如*Pasteurellaceae*科和*Nesseriae*属中的USS序列[20, 23, 24]，作为近亲uptake机制的信号。

- 第二，在DNA的二级结构中可能起重要作用，尤其是分布在非编码区域(intergenic)的重复序列，譬如*Sulfolobus*属中的短规则间隔重复序列[25]，是重要的DNA结构蛋白绑定位点；又譬如，*Haemophilus in-*

*fluenzae*中的IDS[19]和*Eschrecia coli*中的REP[2]，都是dyad结构，很可能在DNA中形成hairpin结构，从而调控其上游的基因表达。

- 第三，直接进入编码区(coding)，改变基因表达，譬如*Haemophilus influenzae*中的tetra tandem重复序列[19]，或者*Nesseriae*属的VNTR序列[26]，其重复单元一般不为三的整数倍，从而在改变分数的情况下改变编码区的相位，导致表达不同蛋白；又譬如一些TE和IS序列，以重复序列为重组的位点，导致基因组的局部序列交换，导致一部分基因表达沉默或者苏醒。

这三个方向也只能是对寡核苷酸重复序列功能的不完全归类，还有更多的其生物功能等待发现。

1.4 寡核苷酸重复序列的分类特异性

寡核苷酸重复序列在上节所述的功能之外，还常带有分类特异的信息。普遍来讲，基因组的差异性往往代表着物种的差异性，另一方面寡核苷酸重复序列恰恰又是基因组的重要组成部分，所以他们才是基因组差异性的载体，也就是物种差异的载体。早在全基因组测序成为现实之前，人们就已经开始自觉利用一些分类特异序列，例如利用种特异或者属特异的序列片断作为引物进行AP-PCR，对该种或者该属进行特异性扩增[27]；例如利用Fingerprint技术检测卫星重复序列，使用这些卫星重复序列作为不同菌株的分子标记，从而区分致病性和无毒性的菌株[28]；也有工作利用相当长度的全蛋白质组短肽序列，并扣除背景，再用成分矢量之间的距离构建细菌的分类树，该完全从基因组数据和计算出发的细菌分类法却与实验学家根据一百多年来的实验积累得到的细菌分类手册吻合的相当好[29]。所有这些实验和计算的事实均表明短核苷酸序列携带有细菌的分类特异性信息，如何从中提取并利用这些信息将是一个很有挑战性的课题。

这篇论文中，我们将从计算得到现今已有的所有细菌基因组的一定长度的寡核苷酸重复序列出发，对这些重复序列的特异性，生物功能和应用作了研究。我们首先编写了计算重复序列的程序，我们的程序REPFINDER能精确完全的得到一定长度的所有细菌基因组中的寡核苷酸重复序列。此外还自己编写了许多辅助工作的Perl程序，来完成中间结果的低复杂度过滤，自重叠过滤等工作。在计算得到中间结果后，我们做了以下分析，首先我们分析了这些计算结果的统计显著性，其次我们在具有统计性的串中寻找了一些特殊的种属或者其他分类级别特异的高度重复寡核苷酸序列并对他们的功能作用做了分析，再次我们搭

建了一个Taxon Specific Oligomer Repeats(TSOR)服务器平台，可以基于用户的需求返回所查询的TSOR序列。这些内容将在下面的论文中都会有详细叙述。

第二章　分类特异性的寡核苷酸重复序列

2.1 研究方法

利用数理统计和K串的组合学及序列算法，我们对所有已测序的细菌全基因组计算其所有的精确重复串。计算过程中我们改变串的长度K。对于这样得到的数据集，我们采用几种方法从中挖掘分类特异的信息。首先，我们设定某个串的在某个物种中的最少出现次数D，同时横向扫瞄其他所有基因组，将这个串在那些基因组中至少出现D次的记录在案。这样我们就能得到在某种显著性条件下，这个串在所有全基因组里的分布状况。其次，我们可以手工选择某一短长度寡核苷酸序列K=12作为种子，从这个种子开始沿展出K=15,18...等含有这个种子串的序列，并对所有这些序列在D=2的情况下横向扫瞄所有基因组，就可以得到关于这个串如何从广泛分布于一些基因组中到，他的衍生串逐渐收缩分布到某些特异的菌株中的分类特异图像。

2.2 数据准备

(1)从NCBI的细菌基因组ftp网站[30]上我们下载了现已测序的所有细菌全基因组的序列文件。所有序列文件的列表见TSOR网站[31]。其中共有293个全基因组，包括古细菌3门10纲12目13科16属24种24个菌株，真细菌14门21纲56目79科120属196种293个菌株。(2)自己设计算法，数据结构并编写程序，计算所有这293个细菌的精确重复的寡核苷酸序列。在计算中，我们需要两个参数，寡核苷酸序列长度K，重复的显著性阈值D。同时我们在计算时选取适当的K的取值范围和步进，实际中我们选择K的范围为9至24，步进为3。在最严格条件下，我们选择D为2，这样经过计算我们就获得了K=9,12,...,24时所有的在细菌基因组里重复的寡核苷酸序列的集合。(3)基于步骤(2)所获得的结果，我们就可以继续研究串的分类特异性了。

2.3 重复串的统计显著性

为了研究寡核苷酸重复序列，寡核苷酸重复序列的统计显著性必须被评估，以使他们区分于随机背景。我们选择了一个二阶的Markov模型来作为我们的

背景。做这样的选择的第一个考虑是，一般而言，我们需要选择一个$3n+2$阶的Markov模型来做背景。这是因为，细菌基因组中的主要部分全部是编码片断(典型情况下，编码区大于90%)。考虑阶数为$3n+2$的Markov模型可以保留$n+1$阶的密码子偏好性关联。这样做的另一个考虑是，如果我们选择一个比二高的$3n+2$阶Markov模型，譬如说5阶，或更高，这样会默许一个假设，在自然界中这些稍长的寡核苷酸片断，5mer或者其它事同等的一致的分布可用的。而实际的观察却并非如此[22]。

我们使用了SSPATT软件来计算统计显著性[32]。Z-score打分机制被用来对短串打分。Z-score的定义如下式：

$$Z-score_{(p)}=\frac{N_p-E_p}{\sqrt{Var_p}}$$

其中N_p是观察到的短串重复次数，E_p是从背景模型期望的重复次

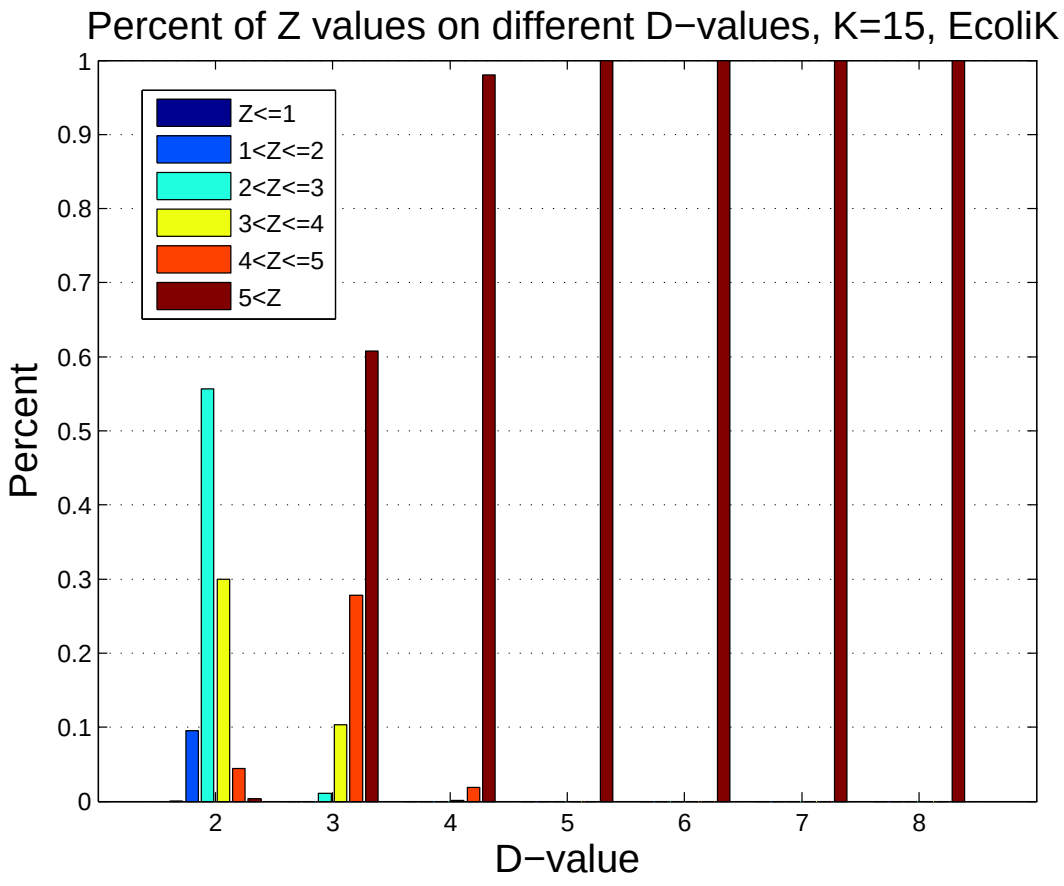


图 2.1 EcoliK, NC_000907, 基因组4.6Mb, GC含量51%, K=15时计算结果

数，Var_p是方差。实际上，Z-score经常用它的对数模式给出。此时，如果一个串的Z-score比3.29要高，那么就意味着只有0.1%的概率这个观察是一个随机事件而不是特殊事件。对于统计显著性直接影响最大的是一个串的重复次数，C值。当串的长度K一定时，一般而言，越大的C值，串的统计显著性越高。除了C值，另一个影响统计显著性的量是串长K。同样的C值下，一般而言，越长的串其统计显著性越高。最后，还有另一个影响统计显著性的量，既是细菌基因组的GC含量偏好性。一个GC含量偏好性很高的基因组相比一个GC含量比较均匀的基因组而言，会使同样一个串的统计显著性减弱。

在这里，我们只需要给出一个比较典型的统计显著性与C值，K值和GC含量偏好性的关系，我们并无必要对所有基因组中的所有寡核苷酸重复串做评估。相反，我们仔细选择了两个典型的基因组来例证C值，K值，GC含量偏好性与统计显著性的关系。我们给出了来自*Escherichia coli* O157:H7 EDL933(EcoliK, Nc_000913, 4.3MB, GC=50%) 的和来自*Xanthomonas oryzae* KACC10331 (Xanor,

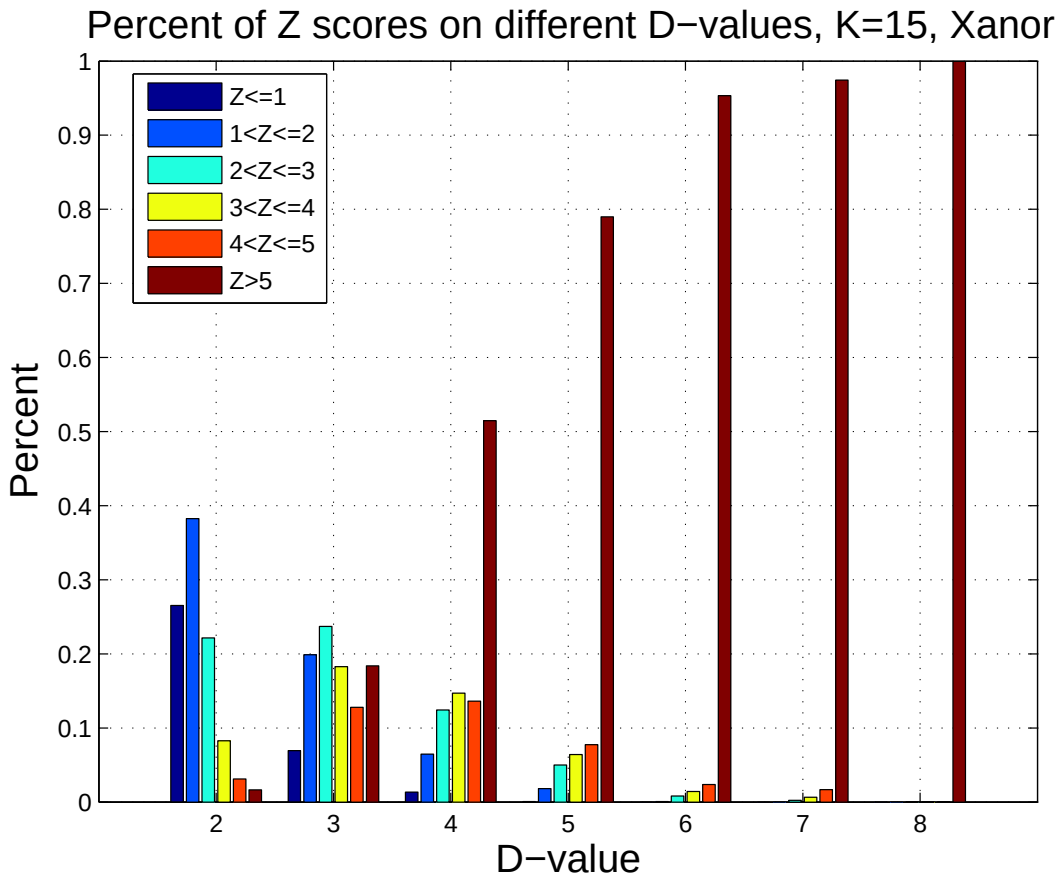


图 2.2　Xanor, NC_006834, 基因组4.8Mb, GC含量68%, K=15时计算结果

NC_006834, 4.8MB, GC=68%)的结果。如下图2.1,2.2,2.3所示，给出了不同Z-score的串占总串数的比例与D值的关系。从图中我们可知，在无GC含量偏好性的EcoliK基因组中，当K=15，C=2时，90%以上的重复寡核苷酸串的Z-score大于3；而在Xanor基因组中,到C=5时,才有90%的寡合苷酸串的Z-score大于3。当K增加到18，此时Xanor基因组中，当C=2即有90%的寡核苷酸重复串的Z-score大于3。因此，如果典型的取阈值Z-score大于3，K=15时，C大于5是保守的统计显著性阈值；当K=18起，保守的讲，任何重复的寡核苷酸短串都具有统计显著性。

2.4 重复串的分类特异性

试图计算得到所有具有分类特异性的串是可能的，但因为计算量庞大，计算不可行也不必要。在现实的研究中，我们更关心那些自身复制份数高的，同时又具有分类特异性的串。如此，我们即可以从某个感兴趣的细菌基因组出

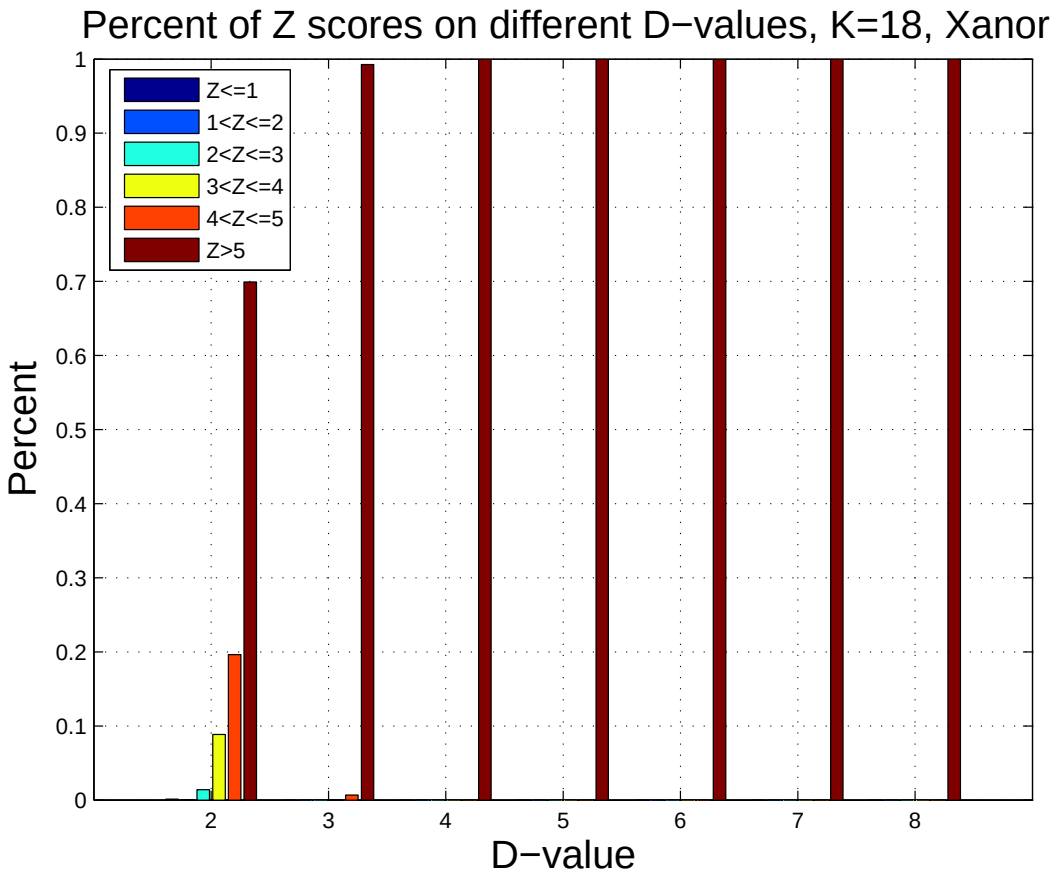


图 2.3 Xanor, NC_006834, 基因组4.8Mb, GC含量68%, K=18时计算结果

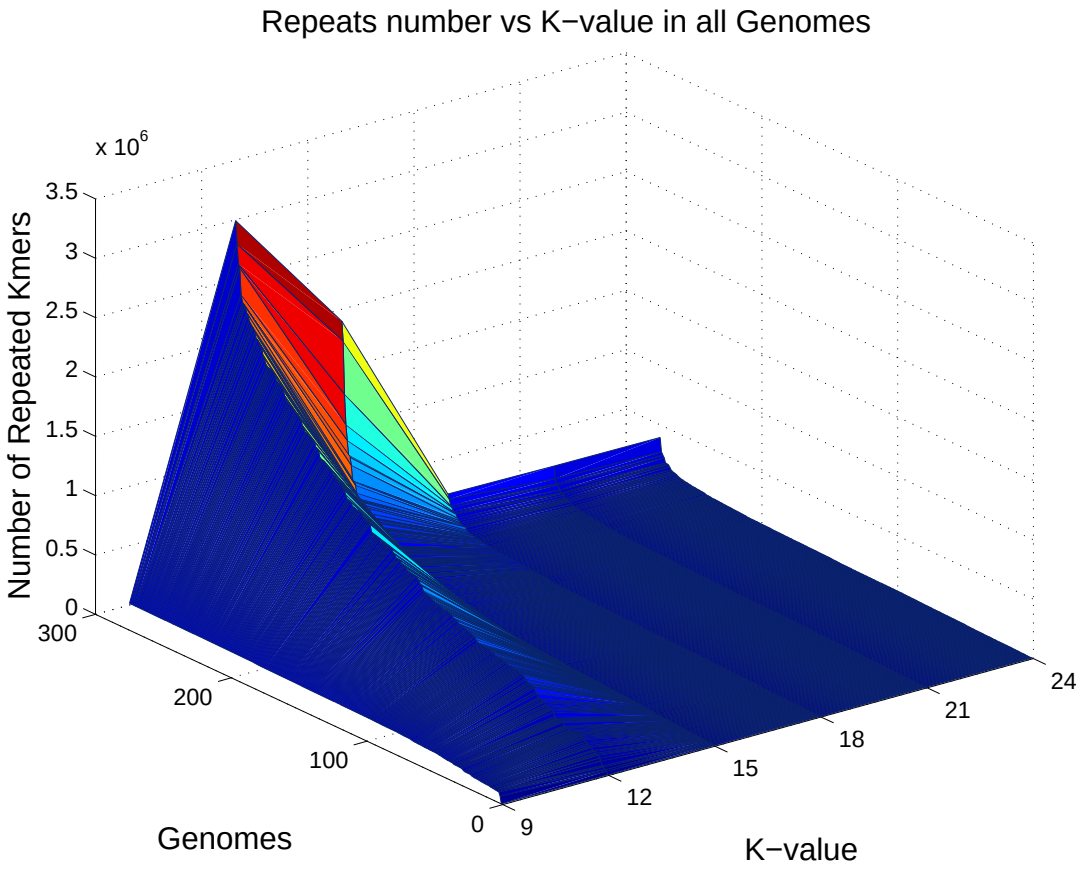


图 2.4　在所有细菌基因组中重复的寡核苷酸串数量和K值关系图：一共包括了275个物种。

发，以某个复制份数为下限阈值，对高于这个阈值的寡核苷酸序列重复序列再来检验其特异性。检验特异性的方法就是检查这个串在其他基因组中的出现次数，在一定长度下，根据统计显著性，我们只需要以重复出现作为标准，如果该串亦重复出现在某一其他基因组中，他既被认为也出现在该基因组中，如果该串只在其他基因组出现一次或者不出现，它既被认为不出现在该其他基因组中。根据这样的出现与不出现定义，我们即可以得到该串出现过的所有细菌基因组，按分类归并这些基因组，我们即可以得到最后这个串所代表的特异分类层。这个层次可以是种，属或者其他的分类层次。

此外，考虑到短核苷酸序列长度K较长时他所代表的分类特异层次会比较高，可以达到种特异或者属特异等等，而在短核苷酸长度K减短时，他所代表的分类特异层次会逐渐降低，会减弱到科特异，目特异等等。可见K值是控制串的分类特异性的最关键参数。直观上讲，越短的寡核苷酸其分布就越广，而越长的就越特异。一个有趣的问题是，在K值的哪个点上，分布的广泛性开始

向特异性转折。试着回答这个问题，我们将K=9，12，...，24的重复寡核苷酸数量给每个物种作了图。将这些图并列在一起，我们得到了下图2.4。从图中可见重复寡核苷酸的数量一般而言在K=12时达到了一个分界点。此分界点前，随着K增大而增多，此分界点后随着K增加而减小，所以特异性增加。由此可见，选择K=12和K=15作为寻找特异性序列的起始K值可行。此外观察随着K长度变化而变化的串的特异性分化过程，我们甚至可以很直观的看到一些基因组的进化线索。很多情况下，这些细节甚至包含着区别细微到菌株之间的一些进化信息。

2.5 一些分类特异的重复串

我们的结果可以从在线的TSOR的web服务器http://tlife.fudan.edu.cn/~xial/tsor.html上访问到（关于TSOR的web服务器的组建和具体使用会在下一章中有详细讲述）。这些结果中，我们有代表性的筛选一些分类特异的寡核苷酸重复序列并对他们进行讨论。

(1)古细菌

- *Pyrococcus* 我们的第一个例子来自于*Pyrococcus*属。*Pyrococci*中共有三个种已被测序，分别是*Pyrococcus furiosus*，*Pyrococcus abyssi*，*Pyrococcus horikoshii*[33]。早先知道在这三个基因组中有一个K=18的串**5'-gttccaataagactaaaa-3'**在各个基因组中均出现，并且重复次数大于20次。而在该三个基因组以外的其它细菌基因组中均严格的不出现，甚至Blast结果表明，该串不出现在其它已知生物的基因组中。所以被认为该串是对于*Pyrococcus*属特异的寡核苷酸短串[34]。在我们的研究中，也发现了该特殊串。

- *Methanobacterium thermoautotrophicus* 和*Archaeoglobus fulgidus*。这是我们第二个例子。这两个基因组共享着一个跨纲特异的K=21的串**5'-aaatcagaccaaatgggatt-3'**，在*Methanobacterium thermoautotrophicus* 基因组中出现了107次而在*Archaeoglobus fulgidus*中出现了171次。这些串高度集中于非编码区域，分别只有0次和19次出现在了编码区。在*Archaeoglobus fulgidus*的测序文章中，这个重复序列被称作LS重复家族[35]。这个重复序列家族与*Sulfolobus solfataricus*和*Sulfolobus tokodaii*中的短规则间隔重复序列(SRSRs)在分布上和结构上有很大相似性，而对SRSRs的研究表明，他们很有可能是属特异的蛋白质的结合位点[25]。由于*Methanobacterium*

thermoautotrophicus 和*Archaeoglobus fulgidus*都是化学自养的喜热菌，这样的重复结构可能在维护基因组在严厉环境下的稳定性有重要作用。

(2)真细菌

- *Escherichia coli*。*Escherichia coli* 是一个有趣的例子。这个种的四个已测序菌株*Escherichia coli* K12，*Escherichia coli* CFT073，*Escherichia coli* O157:H7 和*Eschericia coli* O157:H7 EDL933 从共享特异性寡核苷酸的角度来看，三个野生菌株中的O157:H7 和O157:H7 EDL933与已测序的*Shigella flexneri* 菌株2a str.301 和2a str. 2457T 更接近。而CFT073 菌株则与*Yersinia*属更接近。详细结果可由向TSOR服务器"non-blast"版本递交"Code=B.12.3.13.1, K=15, D=10"查询得到。而经人工选择多年的K12 菌株与其他三个菌株却相距甚远。由它们的测序报告可知[2, 36–39]，K12丢掉了许多转座和移动基因组元素，这可能是之一有趣现象的主要原因。

第三章　寡核苷酸特异序列的发布

3.1 TSOR服务器的设计

设计TSOR(Taxonomic Specific Oligomer Repeats)服务器的目的在于以一种方便用户查询的方式把寡核苷酸分类特异序列的结果共享给用户。整个寻找特异序列的流程如下图所示：

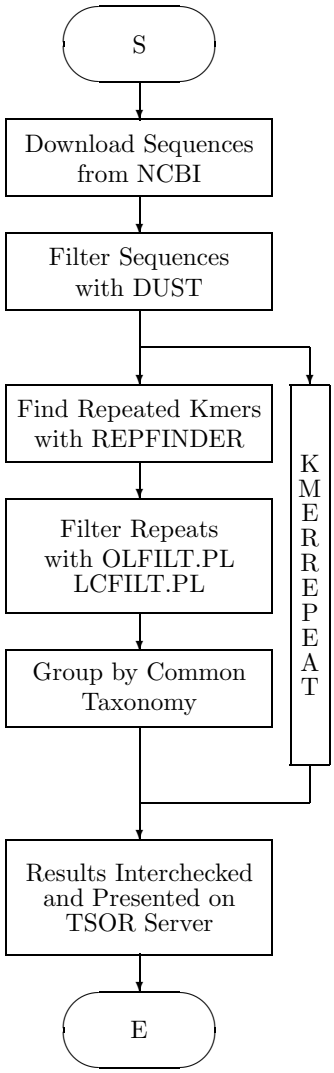
其中步骤的详细操作如下：

(1)基因组数据的准备。如2.2节所述，同时我们使用DUST程序初步过滤下载基因组中的低复杂度重复序列[40]。

(2)寻找寡核苷酸重复序列。利用两个自做的程序，REPFINDER和KMERREPEATER被用来计算过表达的寡核苷酸重复序列，这样结果能互相检验。每次计算过程需要两个参数：K值，计算所寻找的寡核苷酸重复序列的长度和D值，最少的复制份数，大于这个份数的出现将被记录。K值和D值可以是自然数的任意组合，在我们的实际研究中，我们D值的变化从2到以上，不同D值对应的统计显著性已在2.3中讨论。对于K值，我们选择K从9变化到24，步进为3。K值的取舍也在2.3中已讨论。我们使用滑动窗口的方法对寡核苷酸序列进行计数，但这样的方法会对自我交叠的序列会加多计数[41]，所以我们用自己的OLFILT.PL程序来消除前缀和后缀交叠多达一次的重复序列。同时我们再使用另一个程序LCFILT.PL来消除只由一个很小的字母集来生成的序列，例如只有一个或两个字母，以及在头部和尾部有较长单字母重复的序列，例如单字母重复占去序列长度的一半以上者。我们对所有的细菌基因组进行上述操作，这样的结果最终构成我们的中间重复序列数据集。

(3)对寡核苷酸重复序列的分类。我们的TSOR服务器正是基于以上构筑的中间数据集。在TSOR的后台脚本中，我们对于每一次查询，先使用EXKD.PL程序，根据Bergey's Phylogeny Code[42] 找到重复序列最后能归并到的共同分类级别。然后，我们用Local Blast[43]去筛选剔除那些在归并到的分类级别外尚有低重复出现次数出现的那些重复序列。使用Bacteria库作为我们的Blast Database，该库是从所有细菌基因组用Blast的formatDB制作而来的。同时我们定义三个量来表征该寡核苷酸序列的特异性。第一是Self-Hit Number，或叫SH，用来记录来自于同分类层次的hit。另一个叫做Out-Hit Number，或叫OH，用来记录来自

图 3.1　TSOR 流程图


S
Download Sequences from NCBI
Filter Sequences with DUST
Find Repeated Kmers with REPFINDER
Filter Repeats with OLFILT.PL LCFILT.PL
Group by Common Taxonomy
KMERREPEAT
Results Interchecked and Presented on TSOR Server
E

该分类层次外的hit。还有一个叫Outside Starin Number，或叫OSH，用来记录来自于该寡核苷酸序列种子所在菌株基因组之外的Self-hit。

(4)检验结果。刚才提到的另一程序，KMERREPEATER，整合了上述过程，可以用来检验结果。它的不同之处在于它每次只能对某一个基因组做搜索。KMERREPEATER携带三个参数长度K，最低报告SH值sh，最低报告OH值oh。在实际研究中对不同的K，D，sh，oh组合进行了研究，一些例子已经在2.4节中讨论。

3.2 TSOR服务器的使用举例

我们用几个实际的例子来讲述TSOR服务器的使用。

如图3.2所示为向服务器递交查询的用户界面。界面比较简洁，用横线分割为三部分。第一部分是服务器名和当时的版本号。图中所示为”non-blast”的0.1版。第二部分为服务器使用的介绍简单介绍，说明查询递交后，结果会暂时保存在一个网页中，网页的地址会用电邮发给用户。第三部分是服务器的递交表格。首先需要填写用户的电邮地址，其次需要用户填写他所关心的分类层次的Bergey’s代码。该代码可根据菌株名在右侧提供的蓝色链接中找到。此外用户还要设定K值，D值。K值与D值的含义如第2章所言。然后点击提交按钮即可提交任务。第四部分是网站维护者的联系方式。

TSOR服务器现分为两个版本，一是”blast”版，二是”non-blast”版，两者的异同在于：

”non-blast”版，TSOR服务器会返回所有只被用户所提供的Bergey’s代码所代表的分类层次内的基因组所共享的，并且在某个或某些个基因组中重复次数大于等于D值的精确K串。这个版本能帮助用户迅速筛选在被某个分类层次的基因组所共享的高显著性的重复K串。

”blast”版，会将”non-blast”版的结果进一步与细菌库做比对。最后返回的K串会被表明它的OH，SH，和OSH值。这能帮助用户进一步从”non-blast”结果中进一步筛选分类特异的重复K串。表3.1中是用EcoliK作为Bergey’s码，D=10，K=15的从”blast”版本的返回结果的节选，原结果多达364条序列，不便全部在此处给出。EcoliK并不是一个特例，绝大多数我们研究的基因组都带有或多或少这样的菌株特异的寡核苷酸序列。使用KMERREPEAT软件也同样可以得到上述的结果。

有些时候，我们会观察到一些非常显著的特异性重复K串，他们的D值可达

Taxonomic Specific Oligomer Repeats

Taxonomic Specific Oligomer Repeat Server

non-blast version 0.1

Query Submission Form

Use this form to submit your request to TSOR server. TSOR server will discover the TSORs upon your requirement. Results will be presented on a webpage and the address will be sent to you via e-mail.

Your Email Address:

*Provide Taxonomy Bergess Code:*Look for code here

Provide K VALUE *K=9, 12, 15, 18, 21,24*

Provide D VALUE *D>=2, D>20 is recomended*

submit

Please send comments and questions to*lixia at fudan dot edu dot cn*

file:///C|/Documents%20and%20Settings/charlie/Desktop/tsor_query.pl.htm2006-5-14 11:20:40

图 3.2　TSOR服务器查询递交界面

到100以上。表3.2，3.3中我们给出了用”non-blast”版本，K=21，D=100为参数下存在非常显著的特异性重复寡核苷酸序列的基因组。更近一步的”blast”版本显示，绝大多数这些重复串也是严格的具有特异性的。

Pattern	SH	OH	OSH
AGCTGGGTGCCTCAA	11	0	0
AGCCAGCGATTGATG	11	0	0
CTTCATGCCAAAGTG	11	0	0
CAATGCTGCATGCAG	11	0	0
GAGCAGATTCTGCCA	11	0	0
GGCATGAAGGCCCAC	11	0	0
GTCAGTGGGAGAGAT	11	0	0
TAAGCGCCGTCAGAC	11	0	0
TGACGCCTGCTTCGG	11	0	0
TTGAGGCACCCAGCT	11	0	0

表 3.1　EcoliK，D=10，K=15，blast version，TSOR返回结果的节选

Lvl.	Tax.Code	Oligos.	Cp.Num. in Genomes
Cla.	A.2	10	171(Metth),107(Arcfu)
Spe.	A.1.1.4.1.1	2	135(Sulto∗),151(Sulso∗)
Spe.	B.12.2.4.1.1	285	202(Neima∗),191(Neimb∗)
Spe.	B.12.3.13.1.37	1280	107(ShiflT),111(Shifl∗)
Spe.	B.17.1.1.3.1	6	102(Lepin),106(LepinF∗)

表 3.2　共享非常显著的特异性重复K串的微生物基因组

Tax. Code	Abbrev.	Tax. Code	Abbrev.
A.1.1.4.1.1	Sulac	B.12.3.9.1.1	Psepk
B.12.1.6.8.1	Braja	B.12.3.9.1.1	PsepyP
B.12.2.1.3.3	Borpe	B.12.3.9.1.1	PsepyP
B.12.2.6.1.2	Azose	B.12.3.9.1.1	PsepyV
B.12.3.11.1.1	Vibch	B.12.3.9.1.1	Psyar
B.12.3.13.1.26	Pholl	B.12.4.5.2.1	Geosl
B.12.3.3.1.1	Xanor	B.13.1.2.1.11	Thetn
B.12.3.3.1.1	Xancp	B.14.1.1.2.3	Thefu
genomes marked ∗ in table 3.2 also have specific extra. oligo. rep.			

表 3.3　带有菌株特异的寡核苷酸重复序列的微生物基因组

3.3 TSOR序列在实验中的应用

由TSOR得到的分类特异性序列在实验中可能具有应用价值。一种设想是从TSOR的结果中选取一些复制分数较高的特异性序列，同时其在基因组

中的分布又比较均匀，自身又不会互补。然后以这样的序列作为引物，利用ϕ29DNA合成酶对环境基因组进行带偏向性的扩增。在理想状况下可以从环境基因组中特异性的扩增出该分类特异层次的基因组，从而为鉴定环境基因组或者做下一步的全基因组测序奠定基础。目前我们已经和实验组合作，从EcoliK中K=15，D=20，OH小于5的结果中选择了一些引物做了初步试验，结果初步表明具有一定的特异性扩增能力，随着实验的深入，相信会有更大的进展。

第四章　总结

4.1 存在问题及将来方向

在这篇论文中我们用计算的手段研究了细菌基因组中具有分类特异性的寡核苷酸片断。虽然结果比较清晰，但研究中其实还是存在不少问题。

第一，在第2章中，在实际寻找分类特异的重复序列时因为在K=9至12之间重复出现的短核苷酸太多，筛选计算时往往会出现记录空间需求太大，导致内存溢出。在计算寡核苷酸序列的统计显著性时，SSPATT软件在处理列表时会崩溃，必须条条处理，从而效率低下。对于同一基因组的不同K值的寡核苷酸，虽然可以初步看到寡核苷酸从长而专一，到短而特异性变弱，并且随长度变化其分类专一层次减弱的个别现象，但由于受计算能力限制，并且尚无恰当的算法能穷举的在这些要求下找出符合一次性找出符合这些要求的串。在后续的研究中，我们要首先，继续研究算法，设法一次计算获得所有需要的串，其次要找到或自己编写比SSPATT效率更高更稳定的软件，加速统计显著性评估。

第二，在第3章中，虽然TSOR服务器已经初步具备，但现在其功能还相对简易，而且底层基因组数据的更新和维护还需要手工。同时由于服务器计算能力和内存空间的限制，某些K值，D值和Bergey's Code的组合递交后会导致内存溢出，从而计算失败，无法返回结果。再次服务器的使用尚没有详细的在线帮助文档。在后续的研究中，我们会从几个方向着手优化TSOR服务器。首先，修改代码，改进算法和代码效率，更换服务器等，使计算更稳定，解决内存溢出的问题。其次，编写脚本，实现TSOR服务器的自动更新，使其数据库与NCBI的细菌基因组更新同步。

除了以上细节上的更新外，从宏观上，我们希望能引入并行计算，增加TSOR服务器的处理能力，并且向更多人介绍它；寻找实验中的应用可能，譬如上述提到的特异性扩增等；此外还需要结合比较基因组的方法，分析现在已经找到的或者将来找到的分类特异性串的生物功能等。前人的研究及这篇论文中的工作均已佐证一定长度的寡核苷酸序列会是重要的生物信息源和分类信息源，我们有理由相信这个方向上更深入的研究必将带来进化上和整体上对细菌的进一步认识，同时也有可能带给实验学家新的实验研究手段。

参考文献

[1] N.A. Campbell and B.R. Reece. Essential Biology, Chapter14, The Evolution of Microbial Life. Pearson Education, 2001.

[2] F.R. Blattner, G. Plunkett, C.A. Bloch, and et al. The complete genome sequence of *escherichia coli* k-12. Science, 277:1453–1462, 1997.

[3] R.D. Fleischmann, M.D. Adams, O. White, and et al. Whole-genome random sequencing and assembly of *haemophilus influenzae*. Science, 269:496–512, 1995.

[4] Ncbi microbial genome projests, http://www.ncbi.nih.gov/genomes/lproks.cgi.

[5] A.L. Delcher, D. Harmon, S. Kasif, O. White, and Salzberg S.L. Improved microbial gene identification with **glimmer**. Nucleic Acids Research, 27:4636–4641, 1999.

[6] J. Besemer and M. Borodovsky. Genemark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research, 33:W451–454, Web Server Issue 2005.

[7] A. Belkum, S. Scherer, L. Alphen, and H. Verbrugh. Short-sequence dna repeats in prokaryotic genomes. Mcirobiology and Molecular Biology Reviews, 62-2:275–293, June 1998.

[8] H.O. Smith, M.L. Gwinn, and S.L. Salzberg. Dna uptake signal sequences in naturally transformable bacteria. Research Microbiology, 150:603–616, 1999.

[9] E.P.C. Rocha, A. Danchin, and Viari A. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *bacillus substilis* and other competent prokaryotes. Molecular Biology Evolution, 16-9:1219–1230, 1999.

[10] G. Achaz, P. Netter, and E. Coissac. Study of intrachromosomal duplications among the eukaryote genomes. Molecular Biology Evolution, 18-12:2280–2288, 2001.

[11] M.A. Batzer and P.L. Deininger. Alu repeats and human genomic diversity. Nature Review: Genetics, 3:370–379, 2002.

[12] Tigr plant repeats database, http://www.tigr.org/tdb/e2k1/plant.repeats/.

[13] O. Delgrange and E. Rivals. Star: an algorithm to search for tandem approximate repeats. Bioinformatics, 20-16:2812–2820, 2004.

[14] V. Parisi, V. Fonzo, and F. Aluffi-Pentini. String: finding tandem repeats in dna sequences. Bioinformatics, 19-14:1733–1738, 2003.

[15] Gary Benson. Tandem repeats finder: a program to analyze dna sequences. Nucleic Acids Research, 27-2:573–580, 1999.

[16] D. Sharma, B. Issac, G.P.S. Raghava, and R. Ramaswamy. Spectral repeat finder: identification of repetitive sequences using fourier transformation. Bioinformatics, 20-9:1405–1412, 2004.

[17] Akito Taneda. Adplot: detection and visualization of repetitive patterns in complete genomes. Bioinformatics, 20-5:701–708, 2004.

[18] S. Kurtz, V.L. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. Reputer: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research, 29-22:4633–4642, 2001.

[19] S. Karlin, J. Mrazek, and A.M. Campbell. Frequent oligonucleotides and peptides of the *haemophilus influenzae* genome. Nucleic Acids Research, 24-21:4263–4272, 1996.

[20] M. Bakkali, T.Y. Chen, H.C. Lee, and R.J. Redfield. Evolutionary stability of dna uptake signal sequences in the *pasteurellaceae*. Proceedings of National Academy of Sciences, U.S.A., 101-13:4513–4518, 2004.

[21] J. Mrazek, L.H. Gaynon, and S. Karlin. Frequent oligonucleotide motifs in genomes of three streptococci. Nucleic Acids Research, 30:4216–4221, 2002.

[22] E.P.C. Rocha, A. Viari, and A. Danchin. Oligonucleotide bias in *bacillus substilis*: general trends and taxonomic comparisons. Nucleic Acids Research, 26-12:2971–2980, 1998.

[23] K.L. Sisco and H.O. Smith. Sequence-specific dna uptake in *haemophilus* transformation. Proceedings of National Academy of Sciences, U.S.A., 76-2:972–976, 1979.

[24] S.D. Goodman and J.J. Scocca. Identification and arrangement of the dna sequence recognized in specific transformation of *nesseria gonorrhoeae*. Proceedings of National Academy of Sciences, U.S.A., 85-18:6982–6986, 1988.

[25] X. Peng, K. Brugger, B. Shen, and et al. Genus-specific protein binding to the large clusters of dna repeats(short regularly spaces repeats) present in *sulfolobus* genomes. Journal of Bacteriology, 185-8:2410–2417, April 2003.

[26] P. Jordan, L.AS. Snyder, and N.J. Saunders. Diversity in coding tandem repeats in related neisseria spp. BMC Microbiology, 3-23, November 2003.

[27] D.E. Berg, N.S. Akopyants, and D. Kersulyte. Fingerprinting microbial genomes using the rapd or ap-pcr method. Methods in Molecular and Cellular Biology, 5-1:13–24, 1994.

[28] D. Ugarkovic and M. Plohl. Variation in satellite dna profiles - causes and effects. The EMBO Journal, 21-22:5955–5959, 2002.

[29] J. Qi, H. Luo, and B.L. Hao. Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Research, 32:W1–W3, Web Server issue 2004.

[30] Ncbi ftp site for microbial genomes, ftp://ftp.ncbi.nih.gov/genomes/Bacteria.

[31] Taxon specific oligomer repeats web server, http://tlife.fudan.edu.cn/~xial/tsor.html.

[32] H. Richard and G. Nuel. Spa: simple web tool to assess statistical significance of dna patterns. Nucleic Acids Research, 31:3679–3681, 2003.

[33] O. Lecompte, R. Ripp, V. Puzos-Barbe, and et al. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Research, 11-6:981–993, June 2001.

[34] 柏林郝and 淑誉张. 生物信息学手册第二版. 上海科学技术出版社, 2002.

[35] T. Hayashi, K. Makino, M. Ohnishi, and et al. Complete genome sequence of *methanobacterium thermoautotrophicum* dh: functional analysis and comparative genomics. Journal Bacteriology, 179:7135–7155, 1997.

[36] T. Hayashi, K. Makino, M. Ohnishi, and et al. Complete genome sequence of enterohaemorrhagic *eschericia coli* o157:h7 and genomic comparision with a laboratory strain k-12. DNA Research, 8-1:11–22, 2001.

[37] T. Hayashi, K. Makino, M. Ohnishi, and et al. Genome sequence of enterohaemorrhagic *eschericia coli* o157:h7. Nature, 409:529–533, 2001.

[38] R.A. Welch, V. Burland, G.D. Plunkett, and et al. Extensive mosaic structure revealed by the complete genome sequeence of uropathogenic *escherichia coli*. Proceedings of National Academy of Sciences, U.S.A., 99:26–30, 2002.

[39] T. Hayashi, K. Makino, M. Ohnishi, and et al. Complete genome sequence and comparative genomics of *shigella flexneri* serotype 2a strain 2457t. Infection Immunity, 71:2775–2786, 2003.

[40] J.M. Hancock and J.S. Amstrong. Simple34: an improved and enhanced implementation for vax and sun computers of the simple algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Computer Applied Biosciences, 10:67–70, 1994.

[41] L.J. Guibas and A.M. Odlyzko. Periods in strings. Journal Combinatorial Theory, Series A, 30:19–42, 1981.

[42] G.M. Garrity, J.A. Bell, and T.G. Lilburn. Outline: Bergey's manual of Systemetic Bacteriology, 2nd ed. online pulish, 2004.

[43] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. Journal Molecular Biology, 215:403–410, 1990.

发表论文

1. **Li Xia** , Chan Zhou, Phase transition in sequence unique reconstruction Submitted to Journal of System Sciences and Complexity

致　谢

这篇论文是作者从2003年9月至2006年5月间在复旦大学理论生命科学研究中心(Tlife) 研究工作的总结。人生里，要感谢爸爸，妈妈，文戈，以及所有关心爱护我的亲戚，同学，朋友，三年的研究生，七年的复旦就要结束，在你们的支持鼓励下我才能一路走过来。学术上，作者首先要感谢这三年间，郝柏林院士对于研究工作的悉心指导。郝老师不但是我学术上汲取的源泉，更必将以他的言行风度影响我将来的科研道路。此外我还要衷心感谢理论生命科学研究中心的其它老师：谢惠民教授，我和周蝉的第一篇论文来自于他的耐心指导和诚恳建议；郑伟谋教授，与郑老师合作不多，但深为其才子之气折服；袁力老师，喜欢你的八卦和请客。贺平安教授，祝福你在杭州有更好的发展。感谢与我们合作的上海交大赵立平教授，感谢你在生物知识上给与的帮助和其它的一切帮助。感谢周鲁卫教授，与百忙中为我写推荐。当然还有我们实验室的师兄师姐师弟师妹，就不一一列举了，感谢学习生活上你们的帮助，愿你们也毕业顺利，一路好走。