◆ 郭文久 著

微卫星在基因组上的分布与功能及其计算方法初步研究

Preliminary Research on the Microsatellite Distribution and Function in Genomes and the Relevant Computational Methodology

陕西出版集团
陕西科学技术出版社

Preliminary Research on the Microsatellite Distribution and Function in Genomes and the Relevant Computational Methodology

微卫星在基因组上的分布与功能及其计算方法初步研究

郭文久 著

(Wen-Jiu Guo)

陕 西 出 版 集 团

陕西科学技术出版社

内 容 简 介

本书是用计算和统计的方法，系统地研究了基因组微卫星在染色体上的分布、微卫星含量与重组率的相关性、微卫星在水稻基因组各成分上的分布关系、在水稻孤儿与非孤儿基因上的比较基因组学研究结果，微卫星在从病毒、原核到真核物种基因的系统比较基因组研究结果，随后是与此有关的网络计算方法探讨。本书适于从事动植物、微生物及人类遗传育种研究、疾病诊断、法医学等与分子标记基因定位以及分子生态、进化论等领域的研究人员或者爱好者参考。

图书在版编目（CIP)数据

微卫星在基因组上的分布与功能及其计算方法初步研究/郭文久 著
——西安：陕西科学技术出版社，2009.7
ISBN 987-7-5369-4621-7

Ⅰ.微… Ⅱ.郭… Ⅲ.①基因组-研究②分子生物学-研究 Ⅳ.Q343.1 Q7

中国版本图书馆CIP数据核字（2009）第078698号

出版者 陕西出版集团 陕西科学技术出版社
西安北大街131号 邮编 710003
电话（029）87211894 传真（029）87218236
http://www.snstp.com
发行者 陕西出版集团 陕西科学技术出版社
电话（029）87212206 87260001
印 刷 陕西安康天宝有限公司
规 格 880mm × 1230mm 1/16
印 张 14
字 数 380千字
版 次 2009年7月第1版
2009年7月第1次印刷
定 价 38.00元

前 言

分子生物学近年来发展迅猛，尤其是在经过了20世纪末全世界的互联网的快速发展之后，生物信息学借助人类基因组计划的快速实施的东风得到很大的发展。像人类基因组计划于2004年在《Nature》杂志上发表了人类基因组序列的最终版本，似乎人类基因组计划就结束了，然后生命科学的热点就消失了，其实不然，人类基因组计划显然没有像在实施之前想象的那样能够解决所有的包括人类疾病诊断与治疗等在内的生命科学问题。现在盘点人类基因组计划的时候发现，人类基因组计划不但使包括像Lincoln D. Stein等这样的很多科学家因在某一方面的突出成就而一夜成名之外，还为人类留下了一笔巨大的财富——那就是数百块大容量的硬盘都装不下的巨量的分子生物学序列数据。要让这近20年来全世界因为实施人类基因组计划而遗留下来的财富真正造福人类自身，这就是当今生物信息学的重要任务之一。

再来看生物信息学，最近20年来，与通过生物信息学直接得到的结果相比，它所发展的各种算法要多很多。正像《2009年的国家自然科学基金项目指南》中批评的那样，生物信息学可能已经走入了误区，那就是只重视算法的研究而较少重视直接地解决生物学问题本身。实际上，我们翻开任何一期的《Bioinformatics》、《Nucleic Acids Research》、《Genome Research》、《BMC Bioinformatics》等专门发表生物信息学的论文的著名杂志，会发现确实是这样的。这些杂志的影响因子都被抬起来了，作者没有仔细去研究过是算法研究还是直接生物学结果的原因而带来了影响因子的提高。但是一个直观的想法就是，算法距离生物学结果恐怕还有一定的距离。

所幸的是作者的这本专著主要描述的就是用生物信息学的研究方法所得到的具有生物学意义的直接结果。这部专著最早发表于2004年，是作者的博士学位论文。由于论文的某些结果具有一定的科学意义，总是期待能够在SCI影响因子比较高的杂志上发表出来。为了避免与SCI论文原创性的冲突，所以这些内容的中文版本一直不敢出版，期待先发SCI论文，然后再出专著，以便把这些内容链接起来。本书的研究结果经过了在很多SCI杂志上反复投稿反复修改的过程。几经周折，现在终于如愿以偿。

在这本专著中，主要描述了微卫星含量在基因组染色体上的分布、微卫星含量与重组率的各种相关性、微卫星在水稻基因组各个基因成分上的相互关系、微卫星在病毒、原核生物、真核生物基因组上的比较研究，最后是得到这些结果所用到的计算方法描述与体会。属于这部专著的SCI论文只有两篇：一篇是2007年发表在《Comparative and Functional Genomics》，由于是免费杂志，可以自由阅读，位置为：http://www.hindawi.com/getpdf.aspx?doi=10.1155/2007/21676，是关于水稻孤儿基因的研究；另一篇发表在2009年第93卷第4期第323～331页的《Genomics》杂志上，DOI为10.1016/j.ygeno.2008.12.009，这篇就是本专著的核心内容，着丝粒对微卫星含量的抑制。本专著与论文的区别在于在专著上阐明了论文中提及的不成立内容的尝试研究结果以及各个结果之间的相互关系。本书上其他内容也是进一步研究的基础，只是还需要做大量的研究工作。

现在才出版这部专著还有一个很大的原因就是得到了安康学院的学术专著基金的部分资助。同时本专著还部分受到陕西省教育厅自然科学专项基金（项目编号：09JK314）以及安康学院的高层次引进人才的启动项目（项目号：AYQDZR200702）资助才使这部专著得以付印。

目 录

摘　要

微卫星(Microsatellite)是基因组上由1～6个核苷酸为单位组成的重复序列，又称短串联重复序列(Short Tandem Repeat, STR)和简单重复序列(Simple Sequence Repeat, SSR)。在基因组上由于微卫星具有普遍的多态性，是进行群体遗传变异分析、物种起源与进化研究、基因定型(genotyping)、指纹鉴定(fingerprinting)、法医科学(forensic science)、动植物遗传育种等的较好的遗传标记而受到广泛关注。分析微卫星在基因组上的含量、分布及其相关信息，可提高微卫星应用的预见性。特别是目前已有更大的基因组序列数据库，这些研究将获得更新的、更为准确并且具有普遍意义的结果。

本研究采用Perl为编程语言，结合数据库技术等方法，对29个真核生物基因组和1180个原核及病毒生物基因组的DNA序列上长度超过12的微卫星进行了大规模的计算和统计分析。主要探索了如下几个问题：微卫星在染色体上的分布、在基因组范围内的含量和与遗传重组率的相关性、在真核和原核及病毒基因组上分布的共性和个性、微卫星的计算多态性与实验室PCR多态性的关系以及微卫星在孤儿基因和非孤儿基因中的分布性质等。通过这些研究，将为生物遗传、变异和进化发育提供重要的理论和应用基础。此外，还对微卫星的生物信息学本地计算的方法和基于Internet的远程WEB计算方法进行了探讨。取得的主要结果如下：

1. 微卫星在染色体上的分布 在着丝粒及其附近区域的基因组序列中微卫星含量显著低于染色体上的其他区域，远离着丝粒部分的微卫星含量一般都比较高。即使是端着丝粒或者近端着丝粒，其微卫星含量都明显比较低，表现了着丝粒的优先性。在拟南芥、水稻和人类基因组上的计算分析都得到非常一致的结果，但酵母等单细胞真核生物并不遵循这一规律。

2. 在人类和水稻基因组中，计算最近两标记之间的微卫星含量与基因组在此区域的以centi-Morgan(cM)为单位的遗传重组值为数据对组成样本进行了回归相关性分析。计算发现，在水稻基因组中，其spearman秩相关RS统计量等于0.3217，样本容量为2725，其两尾否定概率等于零。在人类基因组上也得到了Spearman RS统计量等于0.1111，样本容量等于2759，其两尾否定概率等于3.9045×10-9。建立了微卫星含量与重组率的线性相关关系。

3. 对29个真核和1180个原核生物基因组中的微卫星进行了计算和统计分析，结果表明：(1)微卫星在真核基因组中的含量一般比原核基因组的高，也有很多物种例外。(2)在真核基因组中微卫星含量在物种之间变异系数不是很大，29个真核物种的平均微卫星含量的变异系数为75%，而以病毒为代表的原核生物的变异系数为91%，推测原核生物是微卫星累积的物种，而真核生物为微卫星含量稳定的物种。(3)微卫星的总motif数量在原核物种中都比较少，在真核物种中一般比较多，那些微卫星含量超过真核物种平均水平的原核物种，其微卫星只是在局部位置的堆积，其motif数量仍然很小。(4)在真核和原核基因组中微卫星的motif数量与基因组大小都成正相关，在原核物种中微卫星含量与基因组大小也成正相关，但在真核基因组中，微卫星的含量与基因组大小不相关。

4. 真核生物的微卫星分布还具有以下特点：(1)微卫星motif在物种之间的使用频率是不一样的，只有a/t微卫星才是所有真核物种所共有的，没有任何一个微卫星motif是某一个物种所特有的。(2)微卫星motif越长，重复次数越少。(3)不同长度的微卫星在基因组内的变异性能是不一样的，长度小于3的微卫星在几乎所有的真核基因组内都有变异，而大于3的则有些变异性非常大，而有的则变异系数等于零；从微卫星总的变异性能看，一般比较长的微卫星变异系数较大。(4)海洋生物的微卫星motif使用频率与陆地动植物有所区别，尤其是长度短于5的微卫星。在海洋生物中，富含c/g的微卫

星在基因组中变异比较大，而在陆地生物中，富含a/t 的微卫星变异较大。

5. 对微卫星在水稻孤儿基因与非孤儿基因之间的关系进行了计算和统计分析，结果表明：水稻孤儿基因总数为28532 条序列，占50.9%，而非孤儿基因为27524 条，占49.1%；孤儿基因中微卫星含量明显高于非孤儿基因；在组成上，不论在孤儿基因中还是在非孤儿基因中，三核苷酸微卫星的含量都超过了50%，孤儿基因中的含量为68%，明显高于非孤儿基因的58%。

6. 对水稻基因组中各成分的微卫星进行了统计分析，结果表明：微卫星在基因的编码区和非编码区都有分布，主要是数量上的差异；在数量关系上，非编码区的微卫星含量大大高于编码区，但是三核苷酸微卫星相反；在水稻EST 中微卫星含量非常丰富；在微卫星motif 长度使用上，二核苷酸微卫星的含量最高，且以at/ta 微卫星占大多数；比较籼稻和粳稻基因组上的微卫星，发现它们在含量、组成和motif 使用频率上都非常相似。

7. 本文对研究中所涉及的计算方法也进行了大量的探索。用Perl 对包括FASTA、GENBANK、XML 和BLAST 报告等的转化和解析进行了编程；还对到TIGR、NCBI 等进行大规模的Internet 远程比对计算进行了编程；使用代理服务器的方法解决了NCBI 只能有50 个同时在线BLAST 的限制；采用Perl Socket 的无阻塞I/O 编程技术，解决了Internet 远程大规模并行BLAST 计算的问题；在程序设计中使用了Perl 的线程技术，极大地提高了网络吞吐量和计算效率，其效率比单线程程序提高了至少10 倍以上。

关键词：微卫星，生物信息学，Perl 编程，着丝粒，真核生物，原核生物，BLAST 基因，基因组，孤儿基因，非孤儿基因，进化论

Abstract

Microsatellite is the tandemly repeated fragment in genomic sequences, composed of 1-6 base pairs in length, so microsatellite also called Short Tandem Repeat (STR) or Simple Repeat Sequence. Microsatellites in genomic sequences are ubiquitously polymorphic, they are broadly applied to population genetic variation analysis, the research on origin and evolution of organisms, genotyping, fingerprinting, forensic science, animal and plant breeding etc.So the analysis of microsatellites about the content, distribution and associations in genomes would provide significant prior information towards microsatellite application, avoiding the scrambles of experimental science. To date with the increase of databases of sequences, the investigation will obtain more precise and novel conclusion.

Using Perl as programming language and integrating databases technology, microsatellites, longer than 12 nucleotides in length, in 29 eukaryotic and 1180 prokaryotic genomes were computed and censused in great scale. The following items were studied: microsatellite distribution on chromosomes, the relationship between microsatellite content and genetic recombination rate in genomic scale, the microsatellite distribution of commonness and individuality in the eukaryotic and prokaryotic genomes, the associations between the polymorphisms in computation and in experiment, the characteristics of microsatellite distribution in orphan and non-orphan genes. The investigation would provide the feasible theories and fundamentals of microsatellite application with significance. Meanwhile the study also explored the bioinformatic computation method in local and remote WEB computation through Internet. The major results showed as follows:

1. Microsatellite distribution on chromosomes showed: microsatellite content in centromeric and pericentromeric regions is notably lower than in others and generally higher in distal regions. Even in acrocentromeric region, the microsatellite content is still lower than chromosomal average, the phenomena expresses the priority of centromere in conservation.The consistent conclusion was confirmed by the computation and statistics of microsatellites in the sequences in Arabidopsis thaliana, Oryza sativa and Homo sapiens genomes, nevertheless the rule could not be concluded in the unicellular prokaryotes such as Schizosaccharomyces pombe.

2. The relationship coefficients between microsatellite content among the closest markers and the recombination rate among the above markers were computed in human and rice genomes. The computation showed the RS statistic of Spearman rank correlation coefficient obtained from the data pairs is 0.3217 with 2725 in sample size in rice (cultivar Nipponbare) genome, while the 2-tailed deniable probability equals to zero, and the same kind of RS statistic, which equals to 0.1111 with 2759 in sample size and the 2-tailed deniable probability also equals 3.9045X10-9, was observed in human genome.

3. The computation and census of microsatellites in 29 eukaryotic and 1180 prokaryotic genomes showed: (1) microsatellite content in eukaryotes is generally higher than in prokaryotes, but with a lot of exceptions. (2) the variation coefficient of microsatellite content in eukaryotes is not so high, which is 75% in 29 eukaryotes, while that in prokaryotes with represent of viruses reaches 91%. So we speculated that the prokaryotes are the microsatellite-cumulative species, while the eukaryotes are the microsatellite-content-stable ones. (3) the total motif number in prokaryotes is lower and that in eukaryotes is higher. Nonetheless the microsatellite content in some species in prokaryotes is higher than in eukaryotes, the microsatellites are only deposited in local in genomic sequences but the motif number is consistently low. (4) That the positive relationship coefficients

between the number of microsatellite motif and genome size in both eukaryotic and prokaryotic genomes were significant, was found. The relationship coefficient between microsatellite content and the genome size was not significant in eukaryotic genomes, but significant in prokaryotic genomes.

4. Other characteristics of the microsatellites in eukaryotic genomes were shown as follow: (1) the genome size did not directly associate with the microsatellite content, but significantly associated with the number of motif. The frequencies of motif are more divergent among species. Only A/T microsatellite is common for all eukaryotes but no motif is unique for a species. (2) The longer of motif, the less of the motif repeated. (3) The variation of different length of motifs within genome is highly diverged. The variation coefficients of all the motif of length less than 3 are above zero, while that longer than 3 are more divergent. Some variation coefficients of motifs are equivalent to zero while others are higher than zero. The trend of the most microsatellite motifs is the longer the motif, the larger the variation coefficient within genome. (4) The frequency of microsatellite motifs in land organisms is differentiated from that in marine organisms and the motif length shorter than 5 is more obvious. In marine organisms, the variation of c/g rich motifs is higher while that of the a/t rich motifs in land organisms is higher.

5. The association of microsatellites in rice (cultivar Nipponbare) orphan and non-orphan genes was computed and censused. The result showed: rice has 28532 orphan genes, which accounts for 50.9%, and 27524 non-orphan genes, which accounts for 49.1%; the microsatellite content in orphan genes is significantly higher than in non-orphan genes; in constitution, tri-nucleotide microsatellite content in ether orphan genes or non-orphan genes exceeds 50%, while the total tri-nucleotide microsatellite content in orphan genes accounts for 68% and that in non-orphan genes accounts for 58%, of total microsatellites.

6. Microsatellites in various components in Oryza sativa SSP. Japanica were computed and censused. The result showed: Microsatellites distributed in both coding regions and non-coding regions of genes and other regions in rice genome. The only quantitative relationship of microsatellites among regions could be confirmed. In quantitative, microsatellite content in non-coding region is higher than in coding region, but the nature of tri-nucleotide microsatellite excepts. That microsatellites in rice EST sequences are plentiful was found similar to the research before. Of the usage of microsatellite motif length, the di-nucleotide microsatellite is higher than others. Of the di-nucleotide microsatellite, the at/ta motif accounts for the major.

7. We also explored the computation methods related to the computation. The format transformation and parse of FASTA, GENBANK, XML and BLAST report were programmed; remote BLAST at TIGR and NCBI though Internet was also programmed; using proxy method overcome the limit of 50 juxtaposition-online BLASTs in concurrence; using multi-thread and non-blocking method solved the issues in parallel remote computation and increased the computation efficiency and communication throughput in comparison with single-thread programming. The efficiency was increased 10 folds at the fewest.

Keywords: microsatellite, bioinformatics, Perl programming, centromere, eukaryote, prokaryote, BLAST, gene, genomes, orphan gene, non-orphan gene, evolution

1 文献综述

1.1 微卫星研究进展

基因组微卫星，又称为简单序列重复(SSR)或者短串联重复(STR)。是指1～6核苷酸碱基的重复序列[1]。现在在几乎所有的基因组上都有发现存在[2,3]，主要是它的重复频率比一般的核苷酸碱基组成要高。Bell[4]认为是由于非偏向的单步随机步行过程造成的，是一种中性序列随机或者近乎随机地分布在常染色质区，Bachtrog等发现at含量与(at/ta)n微卫星密度呈显著正相关，认为微卫星类似随机过程[5]。然而他们发现了有39%的连续序列偏移了微卫星的随机分布。

在最近的文献中反映了有关微卫星进化的相反解释。很多研究都指出，微卫星结构模式受等位基因大小的限制，并且有明显的功能。然而，微卫星在进化上，常常作为一种中性的DNA标记。这些矛盾就更需要微卫星功能的重要证据来综合解释基因组微卫星的进化意义。努力从质量关系上，用功能对中性的观点来分析微卫星的变异现象可能解决不了这个问题。事实上，如果用定量的术语而不是定性术语来解释，微卫星的这种对立解释可能就没有本质上的矛盾了。累计的各种微卫星资料和各种效应证明了这种方法的可行性。

1.1.1 微卫星在基因编码区与非编码区的分布

微卫星构成了非编码DNA相当大的一部分，而在蛋白质编码区则相对的稀少。例如，Wang观察到[6]，在54个植物种中，101个单、二和四核苷酸都位于非编码区。在啤酒酵母(Saccharomyces cerevisiae)、果蝇(Caenorhabditis elegans)、裂殖酵母(Schizosaccharomyces pombe)、家鼠(Mus muscular)、果蝇(Drosophila)、植物和灵长类中[7]都发现微卫星过度(相对于随机)地表现在非编码区中。Morgante等[8]在6种植物(拟南芥、水稻、大豆、玉米和小麦)中除了3、6碱基的微卫星之外，所有的微卫星都存在于编码区中。在日本河豚(Pufferfish)中，只有11.6%存在于编码区中。这是由于在编码区中负向选择对移码突变的矛盾造成的[9]。先前，在三核苷酸微卫星中，有相似的分布模式结果存在于霉菌(Fungi)、原生生物(Protists)、原核生物(Prokaryotes)、病毒(Viruses)、细胞器官(Organelles)、质粒(Plasmids)和人类[10,11]。然而与疾病相关的三核苷酸大多在人基因组的编码区[12]。这与Morgante等的结果相类似，Morgante认为这是由于突变压力和可能的阳性选择特定的单一氨基酸重复的结果。有些三核苷酸重复不是广泛地和长时间地保守的。甚至当它们形成了蛋白质序列的一部分，因为长的三核苷酸重复(如cga重复)能够在减数分裂或者配子发育过程中发生动摇[13]。

在很多物种中主要的微卫星是二聚体核苷酸(48～67%)[6,14]，但是在灵长类中单聚体核苷酸(主要是a/t重复)是最丰富的一类微卫星[11,15]。

对应于三聚体微卫星，二、四聚体的微卫星在编码区就比非编码区少了很多。例如，二聚体核苷酸在挪威云杉中，表达序列比随机的基因组序列少20倍[16]。在8个原核生物和酵母中，几乎所有的长的单聚体和二聚体重复都分布在非转录区[10]。对于在编码区和其他的功能上重要的DNA区域中完全相同的二聚体微卫星，短的、重复单位小于等于3的，能够用Bernoulli模型预测，而在非编码区中长度大于等于5的全二聚体微卫星DNA长度分布适合于无偏单步突变模型[17]。在这个模型中，具有等概率的一次加入或者减去一个重复单元。碱基替换则可以破坏长的全重复，造成两个短的全

重复。在分析人、家鼠、线虫和酵母基因组编码区时，发现所有可能的二核苷酸微卫星分布函数呈指数（Exponential），而同时在非编码区，大多数的二核苷酸微卫星被发现惊人地符合长尾的密定律（Power-Law）函数[18]。这些长而非指数的尾型被推测是非编码区DNA 对突变压力具有较高忍耐力的结果。有很多二聚体微卫星都是分布在基因的5' 端或者3' 端非翻译区。如海峡鲶鱼中的5 个基因[19]，在哺乳类的HSP70 基因具(ga)6cag(tc)24 重复。Lisowska 等也发现二聚体微卫星[20] 在内含子中存在。3' 端和5' 端非翻译区区域和内含子中的二、四聚体微卫星潜在的大小扩增通过移码突变能够导致原基因的破坏或者形成新基因[5, 19]。这个模式认为，如此的二聚体和三聚体核苷酸的随机分布在选择上是相当冲突的[5, 19]。对于同样次数的重复而言，四聚体核苷酸位点就比二聚体位点要长。如果减数分裂的稳定性依赖于目标区域的绝对尺寸的话，这将形成选择压。长重复单位的位点相对于短的位点而言，会经受更大的选择压，在基因组的高重组率区尤其如此[21]。这些发现认为，编码区和非编码区微卫星频率的差异是由于编码区中移码突变造成长度变化之后形成非三联体重复单位的选择压所引起的[18, 19]。

然而，14% 的蛋白质都包含了重复序列，真核生物又比原核生物多了3 倍[22]。 Toth 等[15] 仔细分析了从真菌到人的真核分类中的微卫星，结果显示，在编码区、非编码区，内含子和基因间区域，各种类型的重复形式（从单核苷酸一直到六核苷酸）的分布呈现高度分类特异的模式。这种特异性部分可以用突变机制于不同的选择互作来解释。累计的经验证据似乎显示微卫星序列脊椎动物比无脊椎动物更丰富和更长，而在脊椎动物中，冷血动物的微卫星又要长点。

Eyre-Walker 发现[23] 非编码区的组成变异不能单独地用突变偏向来解释，而选择可能起了重要的作用。与中性突变理论的预测相冲突的是，非编码区DNA 限制在短散布重复的R 带（原始染色质状态）上，而长的散布重复主要是在G 带（Giemsa 黑带）[24] 上。这就让人思考，每个串联重复序列是受制于决定局部不稳定水平和一般的生物活动影响的[25]。非编码区的动态组织形式提供了能够影响密码子使用和稳定染色体的染色质模式的反馈循环[24]。在决定短重复过剩和与长重复相矛盾的选择中受保护的非随机密码子用法，或者整个氨基酸的用法，或者两者都是起了显著作用[10]。等级选择理论显示，选择怎样作用于基因组非编码区，结果创造了位置限制DNA 和形成了个体水平的最小遗传负担[24, 26, 27]。基因组的整个重复水平与基因组的大小和重复程度相关，这样整个基因组可能就反映在与简单重复序列的增加相一致上[15, 28]。这里列出的例子证明了不论基因组内还是物种之间，各种微卫星变异的非随机模式和高度具有功能解释的分类。

1.1.2 微卫星的功能观点

尽管微卫星通常被认为仅仅是进化上的中性DNA 标记，但在各种生物现象中严格地证明了微

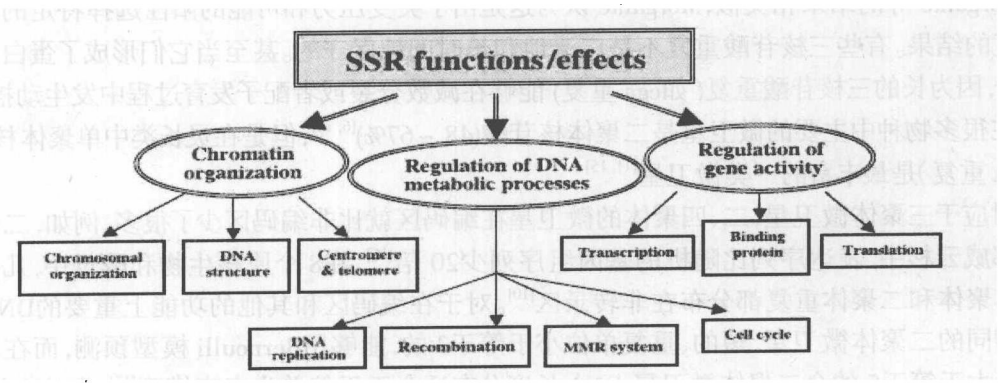

图 1　微卫星的功能与效应 图引自Li等[1]的文献

http://www.researcherid.com/rid/B-4194-2009

卫星具有实质上的功能如图1。

1.1.2.1 染色体组织

某些微卫星分布指出了在分类特异的染色体结构上的可能作用。例如，微卫星的杂交信号与独立于模体（motif）使用的染色体位置相关，并且在普通小麦和黑麦中显示了相当显著的相似分布模式，认为微卫星作为在小麦簇的古代基因组成分上具有特殊作用[29]。位于4A 染色体短臂上的GWM601，ct 重复在野生的二粒小麦维持在(ct)17[1, 30-32]，而在它的后代普通小麦（中国春）中也非常相似[33, 34]，表明这个位点在维持4A 染色体组成的某些方面具有重要作用。这意味着，反复出现的超量微卫星位点，不但在基因组的稳定上起重要作用，在关于附加的基因组特性（如：密码子使用上的进化）上也起着重要作用[10]。

1.1.2.2 DNA高级结构

微卫星 DNA 序列通过简单的和复杂的环- 折叠等可以形成种类繁多的DNA 结构。例如Fragile X (ccg)n 重复形成发夹结构(gaa)n/(ttc)n 形成的双向三重结构，都显示了这种环和折叠。这种三种结构在基因表达上具有重要的调节效应[35]。人着丝粒重复(aatgg)n 能形成双折叠的发夹DNA 结构[36]。相似地，短的三联体重复在单链时也能形成广泛的次级结构[37]。较长的(cag) 和(ctg) 重复在变性和随后的复性中也产生非寻常的次级结构[38]。这种稳定的结构提供了一种解链的机制，这在转录过程中是有利的，也提供给了独特的蛋白质识别模体[36]。在很多物种中，二聚体相对地要丰富些，这表明了基因组序列从随机序列当中分出来，可能也反映了双向曲率、超卷曲和其他高级DNA 结构特征[26, 39]。重复次数似乎是一个严格的参数，这个参数决定了获得在基因表达中所需要的非寻常结构的优势和同样结构在复制时呈现的劣势之间的平衡[36]。

1.1.2.3 端粒与着丝粒

在很多物种中，染色体的着丝粒区域是由无数的串联重复组成，它影响了染色体的组织。在番茄[40]、拟南芥[41]、甜菜[42] 的着丝粒区域存在着长的单、二、三和四核苷酸模体。在（Neurospora crassa）中，着丝粒重复DNA 的基因组Southern 杂交和序列分析显示了包含分化了的着丝粒特定重复的独特着丝粒结构[43]。链孢霉着丝粒简单重复的特性和排列与果蝇的同样成分相似。在果蝇的小染色体上着丝粒邻近DNA 主要包含了高度重复的序列，要求正常传播的重复数在各细胞分裂类型和不同性别上是不同的[44]。

将分化的串联重复序列组装到染色体特异的高度重复上显示很多生物着丝粒所共有的组织特征，这也说明了一种在不同的基因组上创造和维持高度重复的进化机制是保守的[45]。普遍存在于分化物种中的重复序列上以及在主要的收缩位点上提示了在着丝粒结构和功能上之间强烈的进化连接[46]。着丝粒旁临重复DNA 可能起两个作用：姐妹染色体的凝聚和间接协助动粒形成和运转[44]。

1.1.2.4 DNA代谢过程的调节

无数的微卫星和小卫星DNA 被推测可能是重组的热点区域[47, 48]。支持这个思想的实验是猿病毒40[49]、酵母[50]，人[48, 51, 52]、哺乳动物[53]。细菌依赖于RecA 的质粒间重组[54]。其中二聚体核苷酸重复序列是优先位点，因为它们与重组酶具有高度的亲和性[55]。某些微卫星序列直接通过DNA 结构效应影响重组。根据推测，gt、ca、ct、ga、gc 或者at 重复结合蛋白是通过介导的Z 构像或者其他相应的次级DNA 结构参与重组。

重复次数也会影响重组。例如,在体外证实了gt/gc 微卫星重复作用于依赖于RecA 的同源重组的效应。已经发现参与完全链交换的分子数,当(gt) 的重复数包含7、16、37 时,交换率各自从100% 下降到80% 和30%[56]。Majewski 和 Ott 分析了[52] 人22 染色体不同微卫星和重组密度的分布,重组值的增加和微卫星重复仅仅对于GT 重复时是显著相关的。这种效应特别在男性中是显著的。在酵母中,在ARG4 位点的(gt)39 重复显示了增加基因转化的频率,重复序列强烈地刺激多交叉的形成,但对单交叉则无效[57]。这些证据表明,微卫星影响重组不仅仅是通过重复的模体而且还通过重复数。

重组可能潜在地通过非均等交叉或者基因转换改变微卫星的长度。从人的精子中分离的6 个等位点的突变序列的结构分析显示增加位点长度的等位基因间重排的比率和等位基因间的复制事件在等位基因的同源区段趋向于簇集。这两个现象相似于三核苷酸重复不稳定的特征。重组中随机遗传飘移和选择的非均等交换在基因组累积串联重复序列上具有强烈的效应。最近的研究表明了串联重复(包括微卫星和小卫星)不稳定性的非互易重组的重要作用。依赖于模体,非互易的交换可能得到单向的(例如,只有收缩)或者是双向的变化。这种效应要么与减数分裂要么与有丝分裂相关联,虽然其效率不同[57]。

1.1.2.5 DNA复制与细胞循环

微卫星可能影响DNA 复制[10]。在大鼠细胞中,包含d(ga)27.d(tc)27 重复的特定片段捕捕捉到了DNA 扩增。在扩增片段以及与反向重复相关的末端作为在体内DNA 复制的捕捉位点。在哺乳动物突变体表型CSA7 克隆中,不稳定的(ca)n 微卫星重复与其他基因扩增事件是共选择的[58]。

微卫星能够影响控制细胞循环的酶。例如,人CHK1 基因具有控制细胞循环进行的作用。它的编码区包含了(a)9 重复[59],这是一个在肿瘤上具有微卫星稳定性的潜在突变位点[60]。在人结肠和子宫内膜癌中CHK1 基因的改变与稳定的高度多聚A 重复的存在相关联。多聚A 重复中一个A 的插入与删除都将导致截断的蛋白质产生。CHK1 基因的改变[60] 表明了一条癌细胞逃脱细胞循环控制的另一途径。某些控制细胞循环的基因如hMSH6、BAX、IGFIIR、TGFbetaIIR、E2F4 和BRCA2 携带了短重复序列,这些重复序列对于细胞的忠实性和生长控制具有重要意义。微卫星的不稳定性通过插入或者删除一个重复单位影响基因。大多数的微卫星不稳定性肿瘤都是一个基因以上的获得突变,长的重复序列突变的靶标[61]。已有证据表明了DNA 修复与细胞循环的检查点的关联:失配修复系统MMR 与响应(tg)6 或者N- 甲基-N'- 硝基-N- 亚硝基胍介导的DNA 损伤的G2 细胞循环检查点相互作用。在刺小脑无秩序7 型雄性精细胞中发现了非常大的(CAG)n 重复的扩增,如此位点中的很大一部分与胚胎的致命性和功能紊乱的精子相关联[62, 63]。

在真核生物DNA MMR 基因上的微卫星作为进化上的突变调节器,DNA MMR 蛋白质校正了复制错误,在变异的序列中活跃地抑制了重组[64, 65],进而控制了突变率核进化适应性。位于小MMR 基因(MSH3、MSH6、PMS2 和MLH3) 的编码区上的(A)n 座位是真核生物包括人(*Homo sapiens*)、家鼠(*Mus musculus*)、啤酒酵母(*Saccharomyces cerevisiae*)、裂殖酵母(*Schizosaccharomyces pombe*)、果蝇(*Drosophila melanogaster*)、拟南介(*Arabidopsis thaliana*) 和原核的大肠杆菌(E.coli) 的共同特征。尽管在某些物种中,7 碱基单核苷酸重复在主要的MMR 基因(MSH2 或者MLH1) 中只有零星发现,具有指数性质的容易突变的长重复特别容易出现在小的MMR 基因中[66]。微卫星是特别容易产生同步的插入核删除突变,而位于编码区中的非三联体微卫星可能造成高频率的功能突变的移码损失[67]。最近的实验证明对于长的微卫星重复不论在微卫星丰富的家鼠细胞还是在微卫星较少的人类细胞中突变率都是很高的[68]。任意一种小的MMR 蛋白活性的丧失都比主要的MMR 蛋白活性丧失产生较弱的突变表型。失活小的MMR 基因的移码突变率的提高导致真核生物的温和增加突变率的个体向亚群分化。Chang 等提出假说认为[66],在小的MMR 基因中特别高的微卫星含量表示了一种遗传开关,

它容许适应性突变率随着进化时间而调整。

1.1.2.6 基因活性调节

微卫星与转录：很多证据显示当微卫星位于启动子区域时会影响基因的活性。位于启动子区域的(tc)n 重复发现是果蝇[69]、海洋真菌（Aspergillus）[70] 和疫病菌（Phytophthora）[64] 热击蛋白HSP26的转录元素。删除各种二、三和四聚体微卫星重复的后显著地改变了转录活性。例如，从c-KI-ras 基因[71] 和CAT 表达系统的TGF-ß3 基因[72] 的启动子区域删除(tccc)n 重复之后转录活性急剧降低。此外，(gt)n 重复能够提高在方向上距离独立的基因的活性，但最有效的提高转录是在更靠近启动子序列的gt 重复[73]。

位于内含子区域的微卫星仍然能够影响基因转录。例如，位于酪氨酸羟化酶基因的第一个内含子上的四聚体微卫星 HUMTH01 就扮演了转录调节元素的角色[74]。Gebhardt 等[75, 76] 发现位于表皮生长因子(EGFR) 基因第一个内含子上的(ca)n 重复也能影响转录活性。他们也表明RNA 的延长终止与下游靠近微卫星处，这里也是两个分离的转录开始位点。双螺旋DNA 构象的模型计算表明在EGFR多态区域特别是ca 重复较长的区域具有高的弯曲性。这些资料认为，(ca)n 微卫星作为一个连接，它由结合于(ca)n 微卫星的下游的推测阻遏蛋白将启动子带到邻近。值得注意的是，三聚体的微卫星较多地位于与转录和信号转导相关的调节基因中，而编码结构蛋白的基因中则较少，这也表明了在基因转录中微卫星的效应[77]。

基因表达上的重复次数的效应。在很多情况下，微卫星的重复次数显示了对于基因表达和表达水平的关键作用。有些基因只能在微卫星的特定重复次数下表达。例如，位于大肠杆菌lac Z 基因的启动子中的(gaa)12 容许LAC Z 表达，而$(gaa)_{14-16}$ 或者$(gaa)_{5-11}$ 都不允许这个基因表达[78]。有些基因只能在很窄的微卫星重复次数范围内表达，一旦超出这个范围，基因表达就被关闭。在酵母中包含(ctg/cag)n，当n=25 的时候允许URA3 报告基因表达，得到5-Fluoroorotic 酸药物的敏感性，当n ≥30(ctg/cag) 重复次数时URA3 基因就关闭了而出现药物抗性[79]。

其他一组基因表现为对微卫星重复次数的较大宽容。在直接实验帮助测试(tg) 长度对pSV2-CAT（猿猴病毒40 增强子- 加) 或者pA10-CAT（增强子- 减) 表达载体质粒的效应中，最大增强作用是(tg) 在30~40bp 的时候。当(tg) 重复从40 增加到130bp 时。增强子活性降低，130bp 的（tg)重复增强子活性比50bp 时降低了5 倍。有意思的是，poly(tg) 元素的重复范围在人类基因组上位于20 到60bp之间。在这个范围内活性达到最大值[80]。表皮生长因子受体基因的转录活性随着(ca) 重复次数的增加而减少[75, 76, 81]。在携带具有cag 重复的男性荷尔蒙响应元素的CAT 报告系统中，在有二氢睾酮存在时，在25 到77 重复次数的扩增突变中，转录活性随着cag 重复次数的增加而连续降低。在稍微不同的报告系统中也取得了类似的结果，其男性荷尔蒙受体多聚Gln（由多聚cag 编码) 重复长度从0 到50bp[82]。相反的是有些基因的转录水平随着重复次数的增加而增加。例如在人脑PAX-6 基因中，具有(ac)m(ag)n 重复次数≥29 的变异的启动子活性是重复数26 次的4 ~9 倍[83]。在小鸡中构建的(ct) 重复次数从10、15 或者22 的苹果酶基因启动子显示了较野生型(ct)7 增高的表达活性[84]。这些证据表明，在各种生物中有些是自然突变体有些是直接控制实验的结果都显示了在微卫星相关的基因表达调控中微卫星重复次数的重要性。

蛋白质结合实验发现，结合于上游激活序列的有些微卫星是各种调节蛋白的结合位点[85, 86]。例如，单链poly(ga)- 和poly(gt)- 结合蛋白存在于人纤维原细胞中。内含子中的(gt)n 或者混合(gt)n(ga)m 重复在免疫相关基因中至少有70x106 年是保守的，并且对蛋白分子具有很高的亲和性。例如datin 对重复次数为19、15、11bp 的poly(t) 具有相当的亲和性[87]，但就不结合重复次数为3、5、7 或8bp 长的poly(t)。相似的结果也存在于非洲绿猴α- 微卫星结合蛋白对poly(a) 重复的结合[88]。