Computational Studies on Microsatellites in Prokaryotic Genomes with Special Reference to Mycobacterial Genomes

Thesis submitted to

Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad

for the Degree of
DOCTOR OF PHILOSOPHY

Vattipally Bala Sreenu

Registration Number: 2KLBPH15

Centre for DNA Fingerprinting and Diagnostics
Hyderabad
December 2005

University of Hyderabad
School of Life Sciences
Department of Biochemistry
Hyderabad 500 046. India

Declaration

The research work embodied in this thesis entitled, **Computational Studies on Microsatellites in Prokaryotic Genomes with Special Reference to Mycobacterial Genomes** has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of **Dr. Seyed E. Hasnain**. I hereby declare that this work is original and has not been submitted in part or full for any other degree or diploma of any other university.

Vattipally Bala Sreenu

University of Hyderabad
School of Life Sciences
Department of Biochemistry
Hyderabad 500 046. India

Certificate

This is to certify that this thesis entitled, **Computational Studies on Microsatellites in Prokaryotic Genomes with Special Reference to Mycobacterial Genomes** submitted by **Mr. Vattipally Bala Sreenu** for the degree of **Doctor of Philosophy** to the University of Hyderabad is based on the work carried out by him at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted for any diploma or degree of any other university or institution.

Dr. Seyed E. Hasnain
Thesis supervisor
CDFD, Hyderabad

Head, Department of Biochemistry
University of Hyderabad

Dean, School of Life Sciences
University of Hyderabad

Acknowledgements

I am very thankful to Dr. Seyed E. Hasnain, Director, CDFD for accepting me as his student and for his encouragement and motivation throughout my stay at CDFD.

I have thoroughly enjoyed the research that has gone into this thesis. That enjoyment is largely a result of the interaction that I have had with my supervisor and lab-mates. I feel very privileged to have worked with my academic supervisor, Dr. H. A. Nagarajaram, who has transformed me from a microbiologist to a computational biologist, turning my enthusiasm into a career. I owe a great debt of gratitude for his inspiration, philosophical motivation, and friendly guidence. He taught me a great deal about Computational Biology and shared with me the joy of investigation, which is the heart of research.

I would also like to thank Dr. J. Nagaraju, who provided timely support and allowed me to contribute to some of his group's research work especially the development of SilkSatDB.

I would like to thank one of our collaborators, Dr. Ahmed Kamal, Indian Institute of Chemical Technology (IICT), for giving me an opportunity to carryout modelling of DNA-PBD interactions. This work provided me a 'break' in mundane research activities to think and work in a different domain. The collaboration has especially, been very productive to broaden my research outlook.

My thanks to Vishwanath Alevoor and Sushma for setting up microsatellites database (MICdb) and devoloping the primer design program (AUTOPRIMER), respectively. My special thanks to Ranjitkumar Gundu and Priya Sasidharan for their invaluable assistance in the upgradation of MICdb.

My special gratitude to Swetha for her inspiration, support and remote assistance.

I spent many enjoyable moments with my lab members and fellow students chatting about my crazy philosophical ideas over a cup of tea. I thank everyone in my lab, Sridhar, Pankaj, Tabrez, Vishal, Anwar and Suresh for providing a healthy atmosphere and necessary help.

My thanks to my room-mates, Yashin, Rajendra, Sobhan, Subbaiah for making my stay memorable. I also thank Bhaskar and Aravind for their remote help.

I would like to thank Ranjit, Prashanthi, Pavan, Geeta, Swamy and Phani for sharing their knowledge, allowing me to experiment with computers and making my stay in the lab very comfortable. I would also wish to thank Senthil kumar for helping me in typesetting my thesis in LaTeX2$_\epsilon$.

Thanks also to my family who have been extremely understanding and supportive of my studies. Financial assistance received from CSIR, Govt. of India in the form of JRF and SRF is duly acknowledged.

VB Sreenu

Preface

Variation within a population is one of the ways of nature creating biodiversity to ensure survival of the fittest. Heterogeneity provides a high probability for some organisms of a population to be successful in a new environment. Most pathogenic bacteria encounter hostile environments where they have to gain access to host, colonize and proliferate after subverting the host defense system. Thus, it becomes necessary for pathogenic bacteria to maintain high degree of genetic plasticity to confront selections. Mutations in the bacteria would readily create rare variants thereby yielding sufficient genetic plasticity to tolerate difficult conditions. Of the genomic features, microsatellites are known to provide crucial genetic rearrangements with their reversible frameshift mutations which can confer selective advantage.

Microsatellites are tandemly repeated nucleotide sequences of repeating motifs 1-6bp. They are ubiquitously present in all the genomes ranging from viruses and prokaryotes to eukaryotes. Regions with high occurrence of microsatellites, have been termed as contingency loci. One of the unique properties of these repetitive elements is their length polymorphism. They undergo mutations in the form of insertions and deletions (INDELS) of their repeat units, which are in the range of 10^{-6} to 10^{-2} per generation. Microsatellite mutations affect expression levels, cause switching off/on genes and also gene functions.

The basic mechanism behind INDELS of microsatellites is thought to be strand slippage during DNA replication. Misalignments of either the leading strand or lagging strand during DNA replication leads to either insertion or deletion of motifs in the microsatellite tract respectively. These replication errors are usually corrected in the cell by the post replicative mismatch repair system which

consists of genes *mutS, mutL* and *mutH*. Though these genes are conserved across the genera, species like mycobacteria are devoid of these DNA repair genes.

The work presented in this thesis forms a part of an on-going project on systematic studies on microsatellites in prokaryotic genomes involving both computational and experimental approaches. The present work focuses on mycobacterial genomes owing to the fact that these genomes lack the enzymes *mutL, mutS* and *mutH* that constitute the post-replicative DNA repair systems. Hence, as a null hypothesis, it is expected that microsatellite mutations are unregulated. Since mutations are biased towards expansions one can expect long tracts to be present in the mycobacterial genomes. The studies conducted include detailed survey of these genomes for microsatellite distribution, abundance, polymorphism and evolution.

The thesis begins with an introductory review on microsatellites that forms Chapter 1. This chapter details the genesis of the term microsatellite and the discovery of repetitive elements in the DNA in addition to a review of literature available on distribution and frequency of microsatellites in eukaryotic and prokaryotic genomes. The chapter also gives a concise summary of the important properties of microsatellites viz., mutations and their characteristics and their biological consequences. The chapter ends with a note on the experimental and computational methods which were available for microsatellites extractions at the start of the research work.

Although some computational tools were available for the extraction of microsatellites from DNA sequences at the beginning of our studies, we felt the requirement for development of a new tool as none of the existing tools available were able to give automatically information related to the location of microsatellites with respect to coding and non-coding regions in the genome. Furthermore, to our surprise no database was available on prokaryotic microsatellites, on the public domain. These prompted us to develop the program called SSRF and a relational database called MICdb. The details of their development and implementation on the World Wide Web (WWW) form the core of Chapter 2.

Chapter 3 describes the details of the survey of microsatellites in the five mycobacterial genomes *M. avium, M. leprae, M. bovis* and the two strains of *M. tuberculosis* CDC1551 and H37rv. This study was carried out in order to reveal the nature of microsatellite distribution and abundance across the five genomes. These studies were conducted to test the aforementioned null hypothesis concerning any enrichment of microsatellites in light of the absence of the post-replicative DNA repair system. As discussed in this chapter, although these genomes are abundant in short microsatellites, they characteristically show severe scarcity of long microsatellites. This chapter also puts forward an hypothesis on the probable mechanism operating in these pathogens in order to control expansion of microsatellites.

Having studied microsatellite distribution, our interest turned towards the phenomenon of microsatellite polymorphism. Since high quality whole genome sequences of closely related mycobacterial genomes viz., *M. bovis* and the two strains of *M. tuberculosis* CDC1551 and H37rv are available, it was easy to compare these sequences and locate equivalent microsatellites which have undergone length polymorphism. The details of observed microsatellite polymorphisms and their various effects on coding regions are reported in Chapter-4. This chapter also provides an evidence of contingency loci and their involvement in pathogenesis, evolution and adaptation.

As a large number of polymorphic microsatellites observed lead to gene fusion and split events, we specially focused these events in detail and the results of the work are given in Chapter-5.

While the chapters 2-5 deal with studies on perfect microsatellites, Chapter-6 has been devoted to give details of our studies on imperfect microsatellites. Runs of repeat units in a microsatellite get interrupted due to point mutations. Our comparative analysis of the three genomes *M. bovis* and the two strains of *M. tuberculosis* CDC1551 and H37rv reveals interesting effects of point mutations on microsatellites. In addition to arresting microsatellite expansions in some cases, it is revealed that point mutations are also involved in the birth as well as

morphogenesis of microsatellites.

In addition to the studies on microsatellites, we were also involved in studies on simulation studies of DNA with a class of adducts called pyrrolobenzodiazepine (PBD)and the details are presented as Appendix A.

Most of the work presented in this chapter have been either published or communicated and a list of these publications has been provided in the beginning. At the end of each chapter, highlights of the work and conclusions drawn are summarized. The tables and figures referred throughout this thesis are numbered chapter-wise. The references cited in various chapters have been collectively given at the end and are arranged in the alphabetical order. The typesetting of this thesis was done using LaTeX2$_\epsilon$ version 3.3.1.

Contents

List of Publications

1. A. Kamal, N. Laxman, G. Ramesh, P. Ramulu, O. Srinivas, K. Neelima, A. K. Kondapi, **V. B. Sreenu** and H.A.Nagarajaram (2002) Design, synthesis, and evaluation of new non-cross linking pyrrolobenzodiazepine dimers with efficient DNA-binding ability and potent antitumour activity *J. Med. Chem* 45:4679-4688.

2. **V. B. Sreenu**, V. Alevoor, J. Nagaraju and H.A.Nagarajaram (2003) MICdb: Database of Prokaryotic Microsatellites *Nucleic Acids Res.* 31:106-108.

3. **V. B. Sreenu**, G. Ranjitkumar, S. Swaminathan, S. Priya, B. Bose, M. N. Pavan, G. Thanu, J. Nagaraju and H. A. Nagarajaram (2003) MICAS: A fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. *Appl Bioinfo.* 2:165-168.

4. G. Ranjitkumar, M. N. Pavan, B. Bose, S. Swaminathan, G. Thanu, P. Prashanthi, B. P. Prasad, **V. B. Sreenu**, M. S. Achary and H. A. Nagarajaram (2003) EMBnet India Node (EIN) at the Centre for DNA Fingerprinting and Diagnostics: Serving the Indian sub-continent in Bioinformatics. *Bioinformatics India*, 2:67-77.

5. **V. B. Sreenu**, G. Ranjitkumar, S. Swaminathan, S. Priya, B. Bose, M. Narendra Pavan, J. Nagaraju, H. A. Nagarajaram (2004) MICdb - Database of Prokaryotic Microsatellites. *Nucleic Acids Res. Published in online database issue.*

6. M. D. Prasad, M. Muthulakshmi, K. P. Arunkumar, M. Madhu, **V. B. Sreenu**, V. Pavithra, B. Bose, H. A. Nagarajaram, K. Mita, T. Shimada and J. Nagaraju (2005) SilkSatDb: A microsatellite database of the silkworm, *Bombyx mori. Nucleic Acids Res.* 33:403-406.

7. R. K. Gundu, B. Bose, S. Swamynathan, **V. B. Sreenu**, N. Pavan, S. Acharya and H. A. Nagarajaram (2005) Frontiers in Bioinformatics Research: The Biodiversity Issues. Biodiversity : Status and Prospects (Editors: Pramod Tandon, Manju Sharma and Renu Swarup) Narosa Publishing House, New Delhi.

8. **V. B. Sreenu**, P. Kumar, J. Nagaraju and H. A. Nagarajaram (2005) Microsatellite polymorphism across mycobacterial genomes: Implications on genome plasticity and adaptability (communicated).

9. **V. B. Sreenu**, J. Nagaraju and H. A. Nagarajaram (2005) Survey and analysis of microsatellites in mycobacterial genomes (communicated).

10. **V. B. Sreenu**, P. Kumar and H. A. Nagarajaram (2005) Microsatellite length polymorphism causes genes to fuse/split in mycobacterial genomes (communicated).

11. **V. B. Sreenu**, P. Kumar and H. A. Nagarajaram (2005) Evolution of microsatellites through point mutations in *Mycobacterium tuberculosis* and *Mycobacterium bovis* (manuscript under preparation).

Symposia & Poster presentation

- **V. B. Sreenu**, J. Nagaraju and H. A. Nagarajaram (2001) Analysis of microsatellite distribution in prokaryotic genomes. International Symposium on Crystallography and Bioinformatics in Structural Biology, Indian Institute of Sciences, Bangalore, India.

- V. Alevoor, **V. B. Sreenu** and H. A. Nagarajaram (2001) Microsatellite Analysis Server (MICAS). International Symposium on Crystallography and Bioinformatics in Structural Biology, Indian Institute of Sciences, Bangalore, India.

List of Tables

List of Figures

List of Abbreviations

A	Adenine
ATP	Adenosine 5′-triphosphate
BLAST	Basic local alignment search tool
C	Cytosine
DM	Dystrophia myotonica
DNA	Deoxyribonucleic acid
DRPLA	Dentatorubropallidoluysian atrophy
EMBOSS	European molecular biology open software suite
FASTA	Fast alignment
FRDA	Friedreichs ataxia
G	Guanine
HBV	Hepatitis B virus
HD	Huntingtons disease
HGT	Horizontal gene transfer
INDELS	Insertions/deletions
LOS	Lipooligosaccharide
LPS	Lipopolysaccarides
MDR	Multi-drug resistant
MMR	Mismatch repair
MSI	Microsatellite instability
NCBI	National center for biotechnology information
ORF	Open reading frame
PCR	Polymerase chain reaction
SBMA	Spinobulbar muscular atrophy
SCA	Spinocerebellar ataxia
SD	Standard deviation
SMM	Stepwise mutation model
SNP	Single nucleotide polymorphism
SSM	Slipped strand mispairing
SSRs	Simple sequence repeats
STR	Short tandem repeat
T	Thymine
TH	Tyrosine hydroxylase
TRF	Tandem repeat finder
TROLL	Tandem repeat occurrence locator
UTRs	Untranslated regions
VNTR	Variable number of tandem repeat
WWW	World Wide Web
ZNF9	Zinc finger protein 9
aa	Amino acid
bp	Base pair

Chapter 1

Microsatellites: An Introductory Review

1.1 Introduction

The genome sequence of an organism is an array of four nucleotides, Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). However, these four nucleotides are not scattered randomly in the genome. Previous studies including pattern matching (Galas et al. 1985), word frequency counting (Karlin and Burge 1995) and basic linguistic studies (Pevzner et al. 1989a; Pevzner et al. 1989b) have shown the composition of a genome as a non-random distribution of four nucleotides, A, T, G and C.

In every genome that has been sequenced till date, periodicities of nucleotide repeats have been found. Although chance occurrence of a long repeating tract like TGTGTGTGTGTGTG in a typical genome is insignificant, such nucleotide periodicities are exceptionally common in genomes such as the human genome (consortium 2001). Such over-representation of several repeating tracts have been reported (Field and Wills 1998). Of the repeat sequences those constituted by repeating motifs of mono to hexa nucleotides are termed as microsatellites or simple sequence repeats (SSRs). The presence of these abundant repeats is not only limited to the human genome, but also found in all the known genomes of viruses, bacteria and eukaryotes. Since their discovery, the enigmatic over-representation of these repeats has been puzzling researchers.

The term "microsatellite" might appear unusual describing a nucleotide repeat tract, but it carries a little history behind it. Earlier density gradients created by ultracentrifugation of a cesium chloride solution were used in order to study the buoyancy of DNA. When most of the DNA formed a precisely defined layer, a tiny proportion of the DNA formed a satellite band, clearly different from the bulk DNA caused by its lower density (Coreno et al. 1967). Later, the "satellite DNA" was shown to be constituted by repetitive sequences. Since then, researchers started using the term satellite DNA rather than finding a new term to represent repetitive sequences. With further progress, different classes of

"satellite DNA" were defined based on the size of the repetitive element. Direct tandem repeat sequences of motif 10-30bp were called minisatellites (Jeffreys et al. 1985). However, a new class of tandem repeats with shorter repeat motifs were identified and were named as microsatellites following the previous convention (Litt and Luty 1989; Weber and May 1989). Apart from microsatellites, few other types of repeats were also observed in the genomes. These include, palindrome sequences, inverted repeats, and mirror repeats (Cox and Mirkin 1997).

Initial findings of different forms of repetitive sequences in many genomes created an uncertainty in defining the term microsatellite. Although attempts have been made to standardize the nomenclature of microsatellites, there seems to be no consensus among the researchers (Chambers and MacAvoy 2000; Jarne and Lagoda 1996; Tautz 1993). In genomes, chance occurrence of repeats with small tract lengths is more than longer tracts and therefore it is hard to define the minimum iterations needed for a repetitive sequence to be referred to as a microsatellite. For instance, the sequence GCGC occurs frequently in "GC rich" organisms such as *M. tuberculosis*: should this be called as a $(GC)_2$ microsatellite or not? In all the microsatellite analyses reported so far, the threshold assumed for size of the repeat unit and for its number of iterations, seems non-uniform (Ellegren 2004; Tautz 1993). The issue is further complicated with the allowance of imperfection. While some researchers explicitly consider tracts with imperfections (e.g. CACACACA**G**ACACA where at the fifth iteration the C in the repeat motif is substituted by G) as imperfect or interrupted microsatellites (Weber 1990), some others do not make any distinction (Benson 1999) between perfect and imperfect microsatellites. Keeping in view all these we adhere to the definition given by Tautz (1993) according to which a perfect microsatellite is the one characterized by a repeating motif of size within the range 1-6bp (arranged in a head-to-tail manner) iterating at least two times at a given locus, without any imperfection. In the following chapters of this thesis unless otherwise stated we focus our studies only on perfect microsatellites.

1.2 Distribution of microsatellites

Microsatellites are found in all the genomes studied so far. However, with regard to their distribution there is a clear distinction between prokaryotic and eukaryotic genomes. In the following sections we give a brief review of microsatellite distributions in eukaryotes and prokaryotes.

1.2.1 Eukaryotes

Microsatellites are present in all the known eukaryotic genomes. However, the abundance of the different repeat motifs appears to differ extensively depending on the species (Toth et al. 2000). The frequency of microsatellites in eukaryotic genomes is much higher than that could be expected by chance alone (Hancock 1996b). Among the fully sequenced eukaryotic genomes, microsatellites density that is the number of microsatellite tracts in a unit genome, is the highest in mammals, while bird and plant genomes show lower densities (Lagercrantz et al. 1993; Primmer et al. 1997). A comparison of available eukaryotes reveals that microsatellites are most frequently found in *Homo sapiens*, followed by *Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae* and *Cenorhabditis elegans* (Katti et al. 2001). Moreover, these species differ with respect to the abundance of different repeat motifs. The complete sequence of the human genome harbours more than one million microsatellites with the frequency of about one tract per every 2-3 kb and thus constituting about 3% of the genome (consortium 2001). Among these, dinucleotide repeats are more common than mono and tetranucleotide repeats, while trinucleotide repeats are the least abundant. The frequency of the dinucleotide repeats decreases in the order of, CA (50%) > AT (35%) > AG (15%) > GC (0.1%), where CA=AC=TG=GT, AT=TA, GA=AG=CT=TC and GC=CG. In trinucleotide repeats, AAT and AAC, which are about 33% and 21% respectively, are over represented compared to ACC (4%), AGC (2.2%), ACT (1.4%) and ACG (0.1%) (consortium

2001). Comparatively, the mouse genome is repeat-rich consisting of two to threefold more microsatellites than the human genome (consortium 2002). In addition, microsatellite tract lengths are longer in mice than in human and the same is true when the rat and human genomes are compared (consortium 2004). While *Drosophila melanogaster* has shorter microsatellites compared to humans (Bachtrog et al. 1999; Schug et al. 1998), some species of fish have microsatellites that are longer than those in mammalian genomes (Brooker et al. 1994). Long mono and dinucleotide repeats appear to localize in the non-coding regions in many genomes (Field and Wills 1998). In the yeast genome, many of the mono and dinucleotide repeats are AT rich, and the excess trinucleotide repeats are found in the coding regions (Field and Wills 1998).

Microsatellite density is positively correlated to genome size in animals (Hancock 1996a; Katti et al. 2001; Toth et al. 2000). However, in plants, microsatellite density is negatively correlated to genome size (Morgante et al. 2002). The most striking difference between the plant genomes and the other genomes is, the 'AT' repeat is the most commonly found dinucleotide motif in the plant genomes (Lagercrantz et al. 1993; Morgante et al. 2002; Morgante and Olivieri 1993), whereas the same motif is considered as "universally under represented" in all the other genomes (Burge et al. 1992; Karlin et al. 1997).

Microsatellites are distributed more or less uniformly throughout most of the eukaryotic genomes. In the human and mouse genomes, microsatellite density is nearly two-fold higher near the ends of the chromosome arms (consortium 2002). It has also been shown that the distribution of microsatellites on the allosomes differs from that on the autosomes (Bachtrog et al. 1999; consortium 2002). Except for these variations, no other differences have been reported in the density of microsatellites in chromosomes as well as between intergenic regions and introns (Toth et al. 2000).

1.2.2 Prokaryotes

Microsatellites, in prokaryotes as well as in their plasmids, are uniformly distributed (Coenye and Vandamme 2003; Field and Wills 1998; Gur-Arie et al. 2000; Hood et al. 1996; van Belkum et al. 1998; Yang et al. 2003). Compared to eukaryotes, microsatellite tracts are shorter in prokaryotes and they are even shorter in plasmids (Coenye and Vandamme 2003; Yang et al. 2003). In general, short repeats are more prevalent than long tracts (Coenye and Vandamme 2003; Field and Wills 1998; Gur-Arie et al. 2000; Hood et al. 1996). Long tracts comprising of mono and dinucleotide are especially under represented compared to the other tracts, and this under representation is more in the coding regions than non-coding regions (Field and Wills 1998; Gur-Arie et al. 2000). However, some genomes such as *Synechocystis* and *Helicobacter pylori*, show different characteristics in that the observed frequencies match expected frequencies (Field and Wills 1998). In prokaryotes, mononucleotide repeats mostly comprise of A/T motifs (Coenye and Vandamme 2003; Field and Wills 1998; Gur-Arie et al. 2000; Hood et al. 1996). In *Escherichia coli*, di, tetra and pentanucleotide repeats are less than expected and largely comprise of GC/CG repeats. In the *E. coli* genome AT/TA are over-represented in the non-coding regions, while tri and hexanucleotide repeats are over-represented in the coding regions (Gur-Arie et al. 2000). A large number of long tetra-nucleotide repeats is observed in *Haemophilus influenzae* (Hood et al. 1996). In the *Ralstonia solanacearum* genome, di-nucleotide repeats GC/CG are over-represented, while the other di-nucleotide repeats are under-represented, in particular the AT/TA repeats which are significantly underrepresented. The trinucleotide repeats in *R. solanacearum* are over-represented in the coding and non-coding regions. In addition tetranucleotide repeats in this genome, CTAG, AATT, and CATG are under-represented, while those of GTAG and TTAA are over-represented in the chromosome and in the megaplasmid (Coenye and Vandamme 2003). In *Shigella flexneri*, of the total number of microsatellites considered, a higher percentage of them are found in

the coding regions than non-coding regions (Yang et al. 2003).

1.3 Microsatellite polymorphism

Microsatellites exhibit a unique property, which is their high degree of length polymorphism (Weber 1990). Moreover, the length variations are reversible owing to mutations in the form of insertions and deletions (INDELS) of their repeat unit/s, which range from 10^{-6} to 10^{-2} per generation, much higher than base substitution rates (Schlotterer 2000). The degree of polymorphism is both species and locus specific (Amos and Rubinstzein 1996; Ellegren 2000; Harr et al. 1998), but a general tendency towards high degree of polymorphism in long microsatellites has been observed (Brinkmann et al. 1998; Weber 1990). Earlier studies on microsatellite mutations led researchers to propose a length threshold for a microsatellite to exhibit polymorphism (Dechering et al. 1998; Rose and Falush 1998). However, contrasting studies by Pulka and Gruar (1999) showed the non-existence of length limit for microsatellite tracts to show variation, except a positive correlation between the repeat number and the mutation rate. Several other parameters have also been reported to influence microsatellite mutations. A Study by Weber and Wong (1993) showed that the average mutation rate of tetranucleotide repeat tracts is nearly four times higher than that of dinucleotide repeats. However, in a later study, different results were obtained by Chakraborthy et al. (1997), where dinucleotide repeats were reported to have mutation rates of one and a half to two times higher than tetranucleotide repeats. Reports that appeared later introduced further complications by revealing repeat specific mutation rates. For example, in *E. coli*, GT/CA repeats show higher mutation rates than repeats of TC/AG (Eckert and Yan 2000).

1.3.1 Origin of microsatellite length polymorphism

Length variations in microsatellite tracts occur due to the phenomenon known as slipped strands mispairing (SSM) that occurs during replication; also called as DNA polymerase slippage (Levinson and Gutman 1987). The basis of this mechanism was established in the mid 1960s (Fresco and Alberts 1960; Kornberg et al. 1964), and this has been tested and confirmed by *in vitro* experiments (Schlotterer and Tautz 1992). In the course of DNA synthesis, the template strand and the newly synthesized strand temporarily dissociate from each other only to re-associate a fraction of a second later (Figure 1.1). When the newly synthesized strand re-aligns out of phase, renewed replication leads to insertion or deletion of repeat units relative to the template strand (Hile and Eckert 2004.). Most of these primary mutations are corrected by a mismatch repair system (discussed below) and only a small fraction that is not repaired ends up as microsatellite mutations (Strand et al. 1993). A longer repeat stretch supplies larger substrate for the slippage to occur, thus providing an explanation for the length dependent mutation mechanisms seen in several studies (Brinkmann et al. 1998). Replication slippage has been observed and reported in polymerase chain reaction (PCR) experiments (Hauge and Litt 1993; Murray et al. 1993). Moreover, the slippage rate of the Taq polymerase in PCR experiments increases with the number of repeat units and is inversely correlated with the repeat unit length (Shinde et al. 2003).

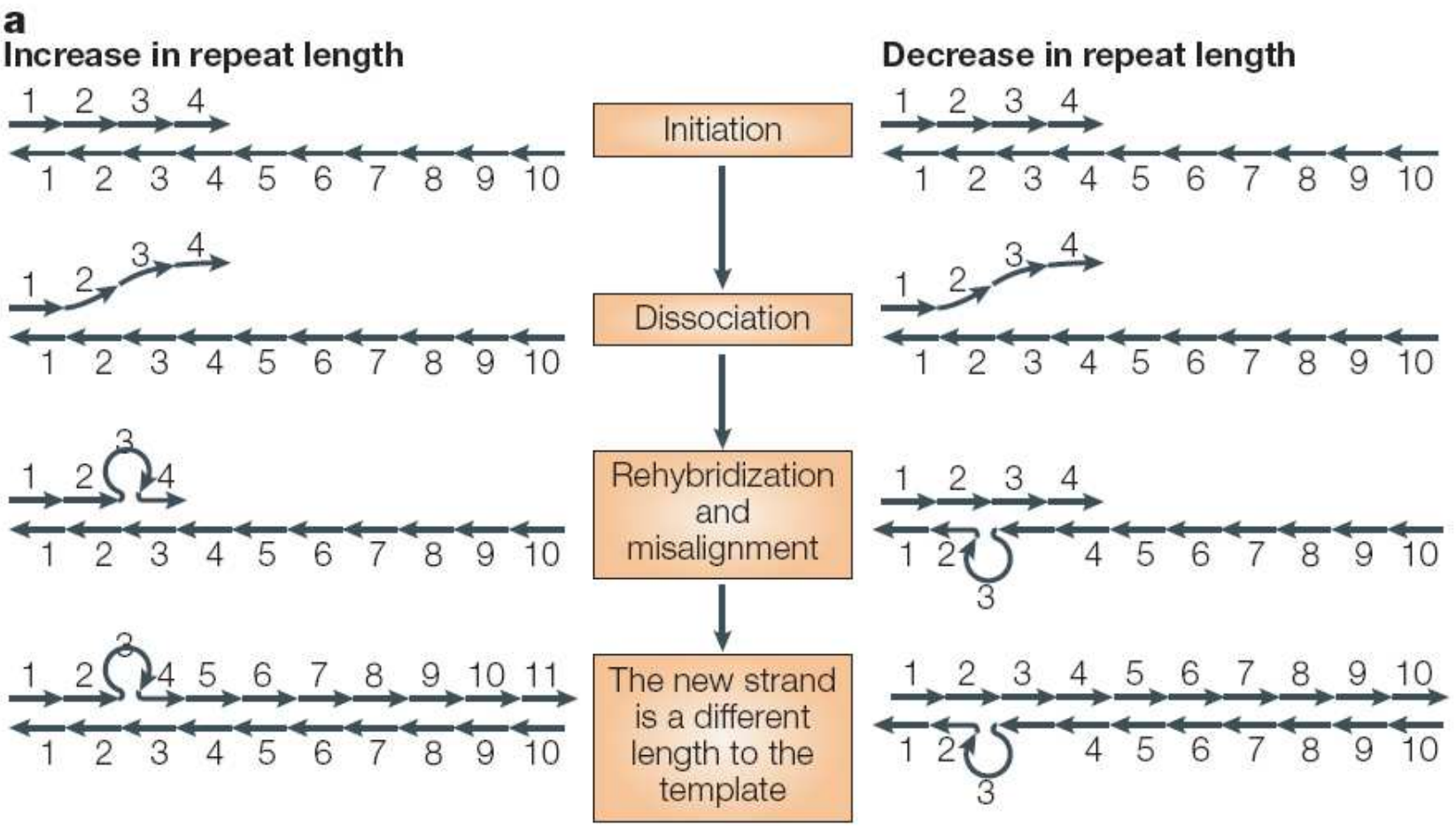

Figure 1.1: An illustration showing the mechanism of slippage strand mispairing during DNA replication

1.3.2 Mismatch repair mechanism

In a cell, microsatellite mediated mutations are corrected by the post replicative DNA repair mechanism also called as mismatch repair (MMR) system. The enzymes involved in the repair process are mutS, mutL and mutH (Acharya et al. 2003). During replication, the adenines (A) in all the 'GATC' sequences of the native DNA strand are methylated by the DNA adenine methylase (Dam), which is a product of the *dam* gene (Calmann et al. 2005). The newly synthesized strand leaves a hemi-methylated DNA state behind the replication fork due to a lag between the replication and methylation by Dam. The MMR system uses the hemi-methylated state of the DNA to differentiate the new strand from the old one for repair. When mismatches arise in the hemi-methylated region of DNA, they are first bound to the MutS protein (Acharya et al. 2003). MutS binds to all mismatches except C-C; it also binds to small insertion or deletion mismatches in which one strand contains one, two, or three extra nucleotides; heteroduplexes with four extra nucleotides are weakly repaired, but larger heterologies do not appear to be recognized (Calmann et al. 2005). MutS then recruits MutL and MutH to form a ternary complex. MutL uses ATP hydrolysis to bring MutS and MutH together, and to stimulate MutH endonuclease activity (Galio et al. 1999). MutH then cleaves the unmethylated strand just $5'$ to the G in the sequence GATC (i.e, N | GATC), leaving a $3'$-OH and $5'$-P at the cleavage site, followed by exonucleolytic digestion to remove the mismatch. The gap produced by the exonuclease action is re-synthesized by the replicative polymerase, DNA polymerase III, which restores the correct nucleotide sequence and the remaining nick is sealed by DNA ligase. Subsequently, the repaired strand is methylated by the Dam methyltransferase at the GATC sequences, thus preventing further action by the MMR system (Calmann et al. 2005). In a typical genome where the repair system is fully operative observed mutations are the end result of the interplay between mutation caused by the slippage and repair mechanisms.

1.3.3 Microsatellite tract constraints

Mutations in a microsatellite tract favour repeat expansions over contractions (Levinson and Gutman 1987). However, a distinct property of microsatellites that has drawn much attention over recent times is their tract length boundary. There are precise length limits for the microsatellite tracts; in the human genome allele lengths exceeding 30 repeat units are rare, and there are always a limited number of these microsatellites in the genome (Bell and Jurka 1997).

1.3.4 Microsatellite mutation models

Step-wise mutation model

Several models have been proposed to explain the dynamic nature of mutations in microsatellites. Most of these models have been derived from the step-wise mutation model (SMM) (Ohta and Kimura 1973). The original SMM suggests that a mutation changes the length of a repetitive array, via the addition or removal of one repeat unit at a fixed rate (Kimmel and Chakraborty 1996; Kimmel et al. 1996; Shriver et al. 1993; Valdes et al. 1993). Later, Di Rienzo et al.(1994) tested the step-wise mutation model for a population with a known demographic history and postulated a modified version of the model called a two phase model that provided a better fit than the original step-wise mutation model. In this model, the vast majority of the mutations are single step mutations but multistep mutations are also allowed at a small rate. Following that, several more complex variants of the step-wise mutation model have also been proposed (Feldman et al. 1997; Garza et al. 1995; Kimmel and Chakraborty 1996; Nauta and Weissing 1996).

Equilibrium model

Since a simple SMM does not explain microsatellite-length distributions (Di Rienzo and al. 1994), other models were proposed to explain the observed variations in the microsatellite tracts. Among these, the equilibrium model is an improved model postulated to explain microsatellite evolution. According to this model, a genome-wide distribution of microsatellite repeat lengths that rests at equilibrium, results from a balance between length and point mutations (Bell and Jurka 1997; Calabrese et al. 2001; Kruglyak et al. 1998). This model proposes three mutational forces that operate on microsatellite sequences. According to this, DNA slippage mutations increase with the increasing repeat count to attain arbitrary high values and random point mutations break these long tracts into smaller units and make them immune to slippage. The random point mutations also create sufficiently long microsatellites which can undergo slippage mutations. In a genome, there is a constant distribution of repeat lengths, governed by the rates of length and point mutations. Later, evidence supporting this model come from a study of homologous microsatellite loci in rat and mouse where long microsatellites were preferentially found in the regions with low substitution rate (Santibez-Koref et al. 2001). This model has been well received in recent years because it explains the differences in microsatellite distribution among species and provides a well-dressed solution to the problem of why microsatellites do not expand into enormous arrays.

1.3.5 Polymorphic microsatellites and their effects

As mentioned earlier microsatellites are highly polymorphic repetitive sequences. Polymorphism of microsatellites that are located far from the coding regions may have very little effect on the gene. However, polymorphic microsatellites residing in the open reading frames (ORFs) and upstream or downstream of regulatory elements produce a considerable effect on the mechanisms of gene transcription as

well as translation and the factors that influence them. Furthermore, the severity of the effect depends on the repeat type and the repeat location. Polymorphic microsatellites of repeating motif length three nucleotides (triplet) or multiples of three always bring out in-frame mutations whereas that of non-triplet repeats bring out either frame-shifts (repeat number variation is not multiple of three) or in-frame mutations (repeat number variation is multiple of three). In the following sections a review on microsatellite polymorphism reported so far and their effects on the various organisms, is given.

1.3.6 Polymorphic microsatellites in human genome

In-frame mutations

coding region: In humans, microsatellite occurrence in genes is controlled by non-perturbation of reading frame. In human cDNA database, more than 92% of the predicted microsatellite variations are caused by microsatellites of length three and multiples of three (Wren et al. 2000). Though, trinucleotides repeat variations result in "in-frame" mutations, expansion of the tract above a certain threshold has been observed to cause disorders, some of which are discussed below. All the known disorders result from triplet expansion to a size that gives rise to pathogenic consequences (Table 1.1). On reaching its threshold level, the size of the repeated array tends to increase with subsequent generations. This property of microsatellites is termed as "dynamic mutations" (Pearson and Sinden 1998; Richards 2001; Richards and Sutherland 1992). This term was coined owing to the size of the array that is unstable from generation to generation, and in some cases, in different cells of an individual.

The triplets repeat expansions in humans are associated with several genetic, neurological and neuromuscular disorders. A majority of these disorders arises due to the microsatellites of the "CAG" repeats. Some of the currently reported disorders of CAG repeat expansions in the coding regions are Huntington's dis-

ease (HD) (Zoghbi and Orr 2000), Dentatorubropallidoluysian atrophy (DRPLA) (Nagafuchi et al. 1994), Spinobulbar muscular atrophy (SBMA) (La Spada et al. 1991) and Spinocerebellar ataxia (SCA, types 1 to 7) (Klement et al. 1998). All of them involve repeat expansions that lead to a single disorder. But expansion of the CAG repeat in the androgen receptor gene leads to various other disorders including abnormal activity of androgen receptor (Coetzee and Irvine 2002), risk of prostate cancer (Buchanan et al. 2001), SBMA with partial androgen insensitivity and is also related to Kennedy's disease (Dejager et al. 2002). Shorter CAG repeats increase the risk of hepatitis B virus (HBV)-related hepatocellular carcinoma (Yu et al. 2002). The only reported disorder caused by non-CAG repeat expansions in human coding regions is oculopharyngeal muscular dystrophy. This is a result of CGC repeat expansion in the PABP2 gene (Brais et al. 1998).

non-coding region: There are other triplet repeat expansions in untranslated regions (UTRs) of the genes, that also cause disorders in humans. Most of these repeats are located in the 5′-UTR of the ORF (Table 1.2). The first genetic disease reported in this category was the fragile X syndrome, the most common form of familial mental retardation (Fu et al. 1991). This syndrome is a result of the expansion of a CGG trinucleotide repeat in the 5′-UTR of the FMRI gene (Fu et al. 1991; Kremer et al. 1991; Verkerk et al. 1991). Soon afterwards, the same loci with different repeat numbers were shown to cause different symptoms. Repeat numbers between 40-60 of the CGG microsatellite are associated with other fragile-X-like phenotypes and woman ovarian dysfunction (Youings et al. 2000). While repeat numbers between 40-200 are related in fragile-X-like cognitive/psychosocial impairment (Franke et al. 1998); a repeat number greater than 200 is reported to result in the loss of FMR-1 function, thus causing mental retardation (Kenneson et al. 2001). A similar repeat expansion observed in the 5′-UTR of the FMR-2 gene causes abnormal neuronal gene regulation (Cummings and Zoghbi 2000).

Other ailments arising from triplet repeat expansions in the UTR regions

include Myotonic dystrophy (DM) (CTG repeat expansion in 3′-UTR of DMPK (Aslanidis et al. 1992)), Spinocerebellar ataxia type 8 (CTG repeat expansion in 3′-UTR of SCA8 (Ranum et al. 1999)), Friedreichs ataxia (GAA repeat expansion in the first intron of FRDA (Campuzano et al. 1996)), and spinocerebellar ataxia type 12 (CAG repeat expansion in 5′-UTR of PPP2R2B (OHearn et al. 2001)).

To date, microsatellite repeat linked disorders are observed only in humans and cannot be remedied after the onset. Study of all these disorders are based only on genetic correlation. A clear involvement of these repeats in the actual disease process is yet to be established. The common feature of all trinucleotide repeat expansion disorders is the existence of a threshold length below which the repeats are normal and beyond which they become "pathogenic". Moreover, this threshold level for the onset of a disease varies among individuals. In all the known disorders, the severity of the disease correlates with the length of the repeat array. However, there is one known exception; in spinocerebellar ataxia type 8, alleles of intermediate size are pathogenic while very long alleles are not (Ranum et al. 1999).

The molecular mechanisms behind the microsatellite expansion process still remain obscure, although the most common explanation is that these motifs create a form of internal hairpin structure when single stranded (Usdin and Grabczyk 2000). So far, only three triplets have been known to be associated with expansion diseases. These are the triplets CGG|CCG, CTG|CAG, and GAA|TTC. The repeats of these triplets can form a variety of non-canonical secondary structures depending on the sequence of the repeat, the number of repeats, the pH, ionic strength, DNA concentration, superhelical density and whether or not the repeat is single stranded (Usdin and Grabczyk 2000). The fact that all hypervariable sequences form secondary structures led to the suggestion that these structures may play a role in the instability of the repeat sequences and in some cases, in disease pathology (Usdin and Grabczyk 2000).

Table 1.1: Microsatellite polymorphism in ORFs leading to in-frame mutations

Orgnism	**Repeat**	**Gene**	**Phenotypic effect due to polymorphism**	**Reference**
Human	CAG	HD	Expansion causes Huntingtons disease (HD)	Zoghbi and Orr (2000)
Human	CAG	DRPVL	Causes dentatorubropallidoluysian atrophy	Nagafuchi et al. (1994)
Human	CAG	KR	Causes spinobulbar muscular atrophy (SBMA)	La Spada (1991)
Human	CAG	SCA	Causes spinocerebellar ataxia (SCA, types 1 to 7)	Klument et al. (1998)
Human	CAG	AR(Androgen receptor)	Shorter repeat increases hepatitis B virus (HBV)related hepatocellular carcinoma risk Modifies prostate cancer risk and progression CAG repeat size links to AR activity Abnormally large AR-CAG sizes result in SBMA with partial androgen insensitivity, which is related to Kennedys disease	Yu et al. (2002) Buchanan et al. (2001) Coetzee and Irvine (2002) Dejager et al. (2002)
Human	CGC	PABP2	Oculopharyngeal muscular dystrophy	Brais et al. (1998)
Drosophila	CAG	DLX6	Triplet expansion leads to cell death	Ferro, et al (2001)
M. hyorhinis	AGT	MG307	Regulates gene translation and influences activity of this surface antigen	Rocha and Blanchard (2002)
E. coli	TCT	ahpC	Converts peroxiredoxin to disulfide reductase	Ritz et al (2001)

Table 1.2: Microsatellite polymorphism in UTRs that is leading to in-frame mutations

Orgnism	**Repeat**	**Location**	**Gene**	**Phenotypic effect due to polymorphism**	**Reference**
Human	CGG	5'-UTR	FMR-1	$(CGG)_{>200}$ cause human mental retardation $(CGG)_{40to200}$ related in fragile-X-like cognitive/psychosocial impairment $(CGG)_{40to60}$ associated in woman ovarian dysfunction	Kenneson et al. (2001) Franke et al. (1998) Youings et al. (2000)
Human	GCC	5'-UTR	FMR-2	Reduced FMR2 causing abnormal neuronal gene regulation	Cummings and Zoghbi (2000)
Human	CAG	5'-UTR	PPP2R2B	$(CAG)_{5578}$ causes SCA12 disease	OHearn et al. (2001)
Human	CTG	3'-UTR	DMPK	Expansion causes DM1 disease	Aslanidis (1992)
Human	CTG	3'-UTR	SCA8	Expansion causes SCA8 disease	Ranum et al. (1999)
Human	GAA	intron	FRDA	Leads to FRDA disease	Campuzano et al. (1996)

Frame-shift mutations

coding region: Microsatellite variations in human genes that are leading to frame-shifts are associated with several cancers and other neuromuscular disorders (Table 1.3). However, all observed repeat changes are not expansions (like triplet repeats to a certain length); variations in the repeat numbers lead to frame-shifts. The first observation of microsatellite instability (MSI) associated with cancer was reported by different research groups in 1993 (Aaltonen et al. 1993; Ionov et al. 1993; Thibodeau et al. 1993). Since then, a wide variety of cancer types have found that are associated with elevated levels of MSI (Halling et al. 1999; Haydon and Jass 2002).

Most of the cancer causing frame-shift mutations due to microsatellite variations are observed in genes involved in the repair mechanism, tumor suppression, cell signaling and cell cycle (Duval and Hamelin 2002; Markowitz et al. 1995; Rampino et al. 1997; Yamada et al. 2002). Majority of the MSI seen in cancers arise due to deficiencies in the primary molecular defense system, i.e the mismatch repair (MMR) system against mutations.

non-coding region: Microsatellite polymorphism in the UTRs especially in the first intron of many genes is reported to cause several diseases (Table 1.5). Variation of the tetrameric repeat of TCAT located in the first intron of the tyrosine hydroxylase (TH) gene is shown to be a regulatory sequence *in vitro* (Meloni et al. 1998). Similarly, the CCTG expansion in the first intron sequence of the zinc finger protein 9 (ZNF9) is involved in the manifestation of Myotonic dystrophy (DM) (Liquori et al. 2001), and polymorphism of the CA repeat located in the first intron of the epidermal growth factor receptor (*egfr*) is associated with breast cancer (Tidow et al. 2003). Intronic changes of poly T in the ATM gene is shown to interfere in splicing and cause colon cancer (Ejima et al. 2000). In spinocerebellar ataxia type 10, large expansions of the ATTCT pentanucleotide repeat is observed in the 9^{th} intron of the SCA10 gene (Matsuura et al. 2000).

Table 1.3: Microsatellite polymorphism in ORFs that is leading to frame-shift mutations in human

Orgnism	**Repeat**	**Gene**	**Phenotypic effect due to polymorphism**	**Reference**
Human	A	hMSH2, hMLH1, hMSH6, hPMS1, hPMS2	Causes human cancers	Vassileva et al. (2002)
Human	A	MBD4/MED1	Causes human cancers	Yamada et al. (2002a)
Human	A	TGFRII, IGFIIR, WISP, GRB-14, AXIN-2	Tumor-suppressive function	Markowitz et al. (1995), Souza et al. (1996)
Human	G	BAX, caspase 5, APAF-1, BCL-10, FAS	Tumor-suppressive function	Rampino et al. (1997), Schwartz et al. (1999)
Human	A	TCF-4, CDX2	Tumor-suppressive function	Duval et al. (1999)
Human	A	2M	Tumor-suppressive function	Bicknell et al. (1996)
Human	A	PTEN, RIZ, Hg4-1	Tumor-suppressive function	Guanti et al. (2000), Zhou et al. (2002)
Human	A	BLM, CHK1, RAD-50	Tumor-suppressive function	Duval and Hamelin (2002)

Table 1.4: Microsatellite polymorphism in ORFs that is leading to frame-shift mutations in bacteria

Orgnism	**Repeat**	**Gene**	**Phenotypic effect due to polymorphism**	**Reference**
H. influenzae	CAAT	*lic1, lic2, lic3*	Affect LPS expression	Roche and Moxon (1995), Hood et al. (1996)
H. influenzae	GCAA	Adhesin	Adhesin phase variation	Hood et al. (1996)
H. influenzae	GACA	Glycosyl transferase	Phase variation	Hood et al. (1996)
H. influenzae	CAAC	Iron binding protein	Phase variation	Hood et al (1996)
H. influenzae	AGTC	Methyl transferase	Host restriction and modification	Hood et al. (1996)
H. influenzae	TTTA	32.9-kDa protein	Unknown function	Hood et al. (1996)
H. somnus	CAAT	LOS component gene	LOS phase variation and adaptation	Inzana et al. (1997)
H. somnus	CAAT	LPS related gene	LPS phase variation and adaptation	Peak et al. (1996)
Neiserria spp	GCAA	Virulence gene	Phase variation and adaptation	Peak et al. (1996)
M. catarrhalis	CAAC	Virulence gene	Phase variation and adaptation	Peak et al. (1996)
N. gonorrhoeae	G	*lsi2*	LOS-specific phenotypic change	Burch et al. (1997)
C. pneumoniae	G	pmp10	Involved in virulence and pathogenesis	Grimwood et al. (2001)
C. pneumoniae	C	Ppp	Involved in the pathogenesis	Rocha et al. (2002)

Table 1.5: Microsatellite polymorphism in UTRs that is leading to frame-shift mutations in eukaryotes

Orgnism	**Repeat**	**Location**	**Gene**	**Effect due to polymorphism**	**Reference**
Human	TCAT	intron	TH gene	Acts as transcription regulatory element	Meloni et al. (1998)
Human	CA	intron	*egfr*	Enhances egfr transcription and involved in breast carcinogenesis	Tidow et al. (2003)
Human	T	intron	ATM gene	Aberrant splicing and abnormal transcription in colon tumor cells	Ejima et al. (2000)
Human	CCTG	intron	ZNF9	Leads to DM2 disease	Liquori et al. (2001)
Human	ATTCT	intron	SCA10 gene	SCA10 disease	Matsuura et al. (2000)
Tilapia (fish)	GT	5'-UTR	*prl 1*	Influence gene expression and growth response of salt-challenged fishes	Streelman and Kocher (2002)

Table 1.6: Microsatellite polymorphism in UTRs that is leading to frame-shift mutations in bacteria

Orgnism	**Repeat**	**Location**	**Gene**	**Effect due to polymorphism**	**Reference**
M. hyorhinis	A	5'-UTR	*vlpA, vlpB, vlpC*	Affect transcription efficiency and are involved in antigenic variation	Yogev et al. (1991)
H. inflenzae	TA	5'-UTR	*hifA* and *hifB*	Influence gene expression and lead to protein adaptation and phase variation	Van Ham et al. (1993)
N. gonorrhoeae	CTCTT	5'-UTR	Outer membrane protein	Regulates gene expression	Murphy et al. (1989)
N. gonorrhoeae	TAAA	5'-UTR	NadA	Regulates the expression of outer membrane protein	Martin et al. (2005)
N. meningitidis	C	5'-UTR	opa	Regulates the expression of outer membrane protein	Sarkari et al (1994)
M. fermentans	A	5'-UTR	P78	Lipoprotein variation	Theiss and Wise (1997)

1.3.7 Polymorphic microsatellites in the other eukaryotes

In-frame mutations

Only a single in-frame mutation in the eukaryotes other than human has been reported which is in the Drosophila homeobox gene, DLX6. In this gene, the CAG repeat expansion results in cell death (Ferro et al. 2001).

Frame-shift mutations

Microsatellite variations are associated with prolactin expression and growth of salt-challenged fish, Tilapia. The repeat variation of GT in the promoter sequence accounts for the regulation of prolactin (Streelman and Kocher 2002).

1.3.8 Polymorphic microsatellites in prokaryotes

Compared to eukaryotes, a lot more information is available on microsatellite repeat variations in prokaryotes. Unlike in eukaryotes, microsatellite variations seemed to have advantageous effects on prokaryotes (Moxon et al. 1994).

In-frame mutations

In *Mycoplasma hyorhinis*, expansion of AGT in the lipoprotein gene (MG307) is shown to regulate gene translation, and thereby influence the activity of this surface antigen (Yogev et al. 1991). One of the radical functional changes observed as a consequence of microsatellite variation is due to the triplet repeat variation in the *E. coli's ahpC* gene. In *E. coli*, the variation of length from $(TCT)_4$ to $(TCT)_5$ in *ahpC* brings about a conversion in the function of the protein, from a peroxiredoxin to a disulfide reductase (Ritz et al. 2001).

Phase variation

Microsatellite polymorphism induced frame-shifts are a major cause of phase variation in prokaryotes, especially pathogenic bacteria. Phase variation as defined by Hallet is "an adaptive process through which bacteria undergo frequent and reversible phenotypic changes resulting in genetic alterations in their genome" (Hallet 2001). Phase variation can be switch on or off in the pili, capsule, flagella or any other protein. The phenotype should be reversible and at a frequency that is greater than that of the spontaneous mutation frequency ($1/10^8$). Phase variation is not necessarily the same as antigenic variation. The on and off states of phase variation can have several genetic mechanisms (Wolfa et al. (in press)). Phase variation was first described in *Salmonella* by Andrewes (Andrewes 1922), where oscillation between the H1 and H2 expression states of flagellar antigens in *Salmonella enterica* serotype typhimurium was shown to vary.

The first report of microsatellite polymorphism affecting expression of a bacterial virulence factor was that ofthe pentanucleotide repeat CTCTT in a gene coding for an outer membrane protein (*Opa*) in *Neisseria gonorrhoeae* (Stern et al. 1986; Stern and Meyer 1987). Subsequently, the role of tetrameric repeats in lipopolysaccharides (LPS) genes was reported in *Haemophilus influenzae* (High et al. 1993; Weiser et al. 1989). There are now a number of examples of microsatellite variations in the translated reading frames or in the promotor regions of many genes from different organisms (Table 1.4 and 1.6). The genes that contain stretches of repetitive sequence and show phase variation due to repeat polymorphism are termed as "contingency loci" (Moxon et al. 1994). Here, we discuss the reported contingency loci and phase variations and their causative results in different organisms.

Phase variations as a result of microsatellite polymorphism in genes, have been extensively studied in *H. influenzae* and *Neisseria* species. In *H. influenzae*, the first genetic loci exhibiting microsatellite polymorphism was *lic1* that is involved in the synthesis of carbohydrate structures on the outer-membrane

lipopolysaccharide (LPS) (Weiser et al. 1989). In the 5′ end of this gene, within the open reading frame, variation of the tetramer repeat CAAT was shown to shift the upstream initiation codons in or out of frame. Subsequently, other genes (*lic2* and *lic3*) consisting of the CAAT repeats and involved in the same pathway were also reported (Robertson and Mayer 1992; Roche and Moxon 1995; Weiser et al. 1990). Sequencing of the *H. influenzae* genome indicated numerous tracts harboring CAAT repeats (Hood et al. 1996). Most of the genes harboring these repeats code for iron binding proteins (Hood et al. 1996). Consequently, similar tetrameric repeats located in genes and displaying phase variations were also identified in other pathogens, like *Neisseria* species, *Moraxella catarrhalis* and *Haemophilus somnus* (Inzana et al. 1997; Peak et al. 1996).

Subsequent studies reported phase variation in *N. gonorrhoeae* and *Chlamydia pneumoniae* brought out by mononucleotide repeat variations in genes. In *N. gonorrhoeae*, polyguanine tract variation in one of the lipooligosaccharide (LOS) synthesis genes, namely, *lsi2* was shown to be responsible for LOS-specific phenotypic variation (Burch et al. 1997). Computational analysis of the complete genome of *N. gonorrohoaea*, indicated a total of 65 potential phase variable genes in the genome (Saunders et al. 2000). The phase variation of the polymorphic membrane proteins (Pmp) in *C. pneumoniae* due to polymorphism of polyguanine tracts, has been shown to be involved in virulence and pathogenesis (Grimwood et al. 2001). Recent *in silico* and *in vitro* analysis have reported polycytosine variation in uncharacterized species-specific proteins called pneumoneae polymorphic protein (ppp) (Rocha et al. 2002). These proteins are further characterized as outer membrane proteins and are involved in the pathogenesis of *C. pneumoniae*.

Microsatellite variations in the upstream regions of genes are also involved in phase variations in pathogenic bacteria (Table 1.6). A contiguous strand of adenine residues to the upstream of the -10 box that is subjected to frequent mutations, has been shown to regulate phase variation of size-variant membrane surface lipoproteins (*Vlps*) in *Mycoplasma hyorhinis* (Yogev et al. 1991). In *H. in-*

flenzae, variation of the TA repeat in the promoter region of *hifA/hifB* genes that encode fimbrial subunit proteins is a classical example of microsatellite involvement in gene regulation (Van Ham et al. 1993; Van Ham et al. 1994). Reversible phase variation is due to changes in the number of TA repeats, that space the 35 and 10 boxes of the dual promoter controlling *hifA* and *hifB* and shown to regulate the fimbrial genes. In *N. gonorrhoeae*, expression of the outer membrane protein P.II is controlled by the pentamer repeat 'CTCTT' in the leader sequence (Murphy et al. 1989). Another gene regulated by microsatellite variation in the promoter sequence is *NadA*, that codes for outer membrane adhesion protein in *N. gonorrohoaea*. Expression of this gene is under the control of the TAAA repeat number (Martin et al. 2005). In the other species of Neisseria like *N. meningitides*, transcription of the outer membrane protein (*Opa*) is under the regulation of the polycytidine repeat tract (Sarkari et al. 1994). Frame-shift mutations in the promoter region of the antigenic membrane protein (P78) are regulated by the variation of ployadenine repeats in *Mycoplasma fermentans* (Theiss and Wise 1997).

Microsatellite mutations and adaptations

In general, the examples mentioned above emphasize the importance and influence of microsatellite elements in many aspects of adaptive bacterial behavior. The significance of these DNA repeats lies in their relative instability through mutational mechanisms that result in an increase or decrease in the number of repeats. This in turn brings about a frame shift when the number of repeats changes within a DNA sequence that is translated or in altered transcription when the change is located within the promoter.Thus, translation or transcription of encoded molecules is switched on or off resulting in phenotypic (phase) variation. This switching is reversible and occurs at high frequencies,typically about 10^{-2} per bacterium per generation (Moxon et al. 1994). Majority of the proteins encoded by these variablegenes are cell surface proteins like adhesions

or invasions, while others are glycosyl transferases involved inthe biosynthesis of LPS. Hence, it appears that the repeats in the genes make them as reservoirs for bringing out certain variability in virulence, antigenicity and host adaptation. It is thought that these mechanisms of phenotypic variations have evolved in order to facilitate bacterial pathogens to adapt to the changing microenvironments within and between hosts. Many infections involve the clonal expansion of a microbial population and phase variation seems to provide a combinatorial mechanism whereby pathogenic bacteria can vary several host interactive surface molecules, reversibly and independently, within orin between hosts.

The role of microsatellite variation is in adaptation of non-phathogenic bacteria is well documented; in *E.coli*, microsatellites respond adequately to conditions of nonlethal selection. Detailed studies focusing on the molecular basis of these phenomena that results in reversible frameshift mutations in *E.coli* have been published (Foster and Trimarchi 1994; Rosenberg et al. 1994). Small variations in a homopolymeric tract were identified upon nutritional deprivation. SSM was proposed as the basic mechanism for this type of adaptive behavior.

1.4 Microsatellite extraction methods

1.4.1 Experimental methods

Traditionally, microsatellite loci have been isolated from partial genomic libraries (selected for small insert size) of species of interest, screening several thousands of clones through colony hybridization with repeat containing probes (Rassmann et al. 1991). Although relatively simple, for especially microsatellite rich genomes, this approach can turn out to be extremely tedious and inefficient for species with low microsatellite frequencies. Therefore, several alternative strategies have been devised in order to reduce the time invested in microsatellite isolation and significantly increase yields (Cifarelli et al. 1995; Fisher et al. 1996; Primmer et

al. 1996; Williams et al. 1990;).

1.4.2 Computational methods

In the post-genomic era, the availability of complete genome sequences has simplified and eased the task of screening genomes for microsatellites. A computer program specially developed for microsatellite screening, is sufficient to scan the whole genome for sequence composition of motifs, locations and frequencies of microsatellite tracts. There are many computational tools available for the extraction of microsatellites from genomic regions.

Tandem repeat finder

This program is designed to find all tandem repeats in the nucleotide sequences (Benson, 1999). No specification of either the pattern or pattern size is required for this program. TRF models tandem repeats by percent identity and frequency of indels between adjacent pattern copies using statistically based recognition criteria. A World Wide Web (WWW) server interface http://atc3.biomath.mssm.edu/trf.html has been established for the automated use of this program. The input required for this program is a sequence in FASTA format, to be submitted by the user. The result of the analysis is sent back to the web browser of the user as two files, a summary table file and an alignment file. The summary table contains information about each repeat, including its location, size, number of copies and nucleotide content. Clicking the location indices of the table entries, results in the opening of a second web browser that provides the alignment of the copies against a consensus pattern. The program is extremely fast, analyzing sequences in the order of 0.5Mb in just a few seconds. Submitted sequences may be up to a maximum of 5Mb in length. Repeats with pattern size in the range of 1 to 500 bases are detected. The sequence information sent to the server is kept confidential and deleted after program execution.

Sputnik

Sputnik is a C language program that searches DNA sequence files in FASTA format for microsatellite repeats. It finds SSRs, while allowing for deviations from a perfect repeat. The output from the Sputnik are amenable for other programs for analysis. When the input sequence is non-redundant, the process involves simply selecting a repeat length and percent perfection level, and subsequently counting the results. This program is available for free download on WWW (http://espressosoftware.com/pages/sputnik.jsp)

Tandem Repeat Occurrence Locator

The Tandem Repeat Occurrence Locator (TROLL), is a light-weight dictionary based simple sequence repeat (SSR) finder based on a slight modification of the Aho-Corasick algorithm (Castelo et al, 2002). It is fast and requires only a standard personal computer (PC) for operation. This program finds repeats of length 20 bp or more on a given sequence. TROLL is an open source project and is available at http://finder.sourceforge.net.

Poly

This is another similar repeat finding program for extracting microsatellites from a given nucleotide sequence input (Bizzaro and Marx, 2003). It has been developed using the Python programming language. In comparison to Sputnik and Troll, Poly requires greater computational runtime for analysis.

1.5 The present work

Though the involvement of SSRs in phase variation and virulence is well documented in most pathogenic bacteria, there are no such reports pertaining to the

highly pathogenic bacterium *Mycobacterium tuberculosis*. *Mycobacterium tuberculosis* is the causative agent of tuberculosis, and is an ancient disease causing bacteria known to mankind. Tuberculosis causes more deaths worldwide than any other infectious disease (Dye et al, 1999). Every year, approximately 2 million people in India develop tuberculosis (TB), accounting for one fourth of the world′s new TB cases (Swaminathan et al, 1999). Added to this, multi-drug resistant strains of *M. tuberculosis* (MDR TB) are increasingly being found in the clinics. It continues to remain a challenge to researchers, to understand the various virulence factors involved in the actual disease manifestation of tuberculosis.

The *Mycobacterium tuberculosis* genome has been studied for repetitive sequences since many years (Hermans et al 1991; Hermans et al 1992; van Soolingen et al 1993; Kamerbeek et al, 1997). However, all these studies mainly focused on insertion elements, that are nowadays used either as markers for detection of *M. tuberculosis* or for epidemiological studies. Although the SSR loci are directly or indirectly involved in the pathogenicity and definition of contingency genes, no efforts have beeen carried out to study contingency genes in this pathogen. This could perhaps be owing to the painstaking process involved in the *in vivo* study of highly mutable genes. In addition, the generation time and fastidious growth requirement of *M. tuberculosis* pose as great hurdles towards *in vivo* studies.

One of the important reasons in undertaking this study of mycobacterial genomes is the interesting feature of these genomes being devoid of the post replicative repair mechanism, that is very essential for microsatellite expansion and stability in the genome.

1.6 Summary

In this chapter an overview of microsatellites in eukaryotic and prokaryotic genomes with a special reference to their distribution, abundance and polymorphisms, has been presented. The important property of the microsatellites is their ability to

undergo insertion/deletion of their repeat units. Mutations of this type affect transcription as well as translation of coding regions at the immediate vicinity of the microsatellite which in turn bring out certain variability to the genome as a whole. As discussed, microsatellite mutations are advantageous to prokaryotes, offer variability needed to survive hostile host environments and may also introduce new varieties.

In the chapters to follow, we present a report on the research work that has been carried out on computational analysis of microsatellites in mycobacterial genomes.

Chapter 2

Development and hosting of MICdb - a relational database of microsatellites extracted from the genome sequences of prokaryotes and viruses - on the world-wide-web

---

Publications from this chapter

- **V. B. Sreenu**, V. Alevoor, J. Nagaraju and H.A.Nagarajaram (2003) MICdb: Database of Prokaryotic Microsatellites *Nucleic Acids Res.* 31:106-108.
- **V. B. Sreenu**, G. Ranjitkumar, S. Swaminathan, S. Priya, B. Bose, M. N. Pavan, G. Thanu, J. Nagaraju and H. A. Nagarajaram (2003) MICAS: A fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. *Appl Bioinfo.* 2:165-168.
- **V. B. Sreenu**, G. Ranjitkumar, S. Swaminathan, S. Priya, B. Bose, M. Narendra Pavan, J. Nagaraju, H. A. Nagarajaram (2005) MICdb - Database of Prokaryotic Microsatellites. *Nucleic Acids Res. Published in online database issue.*

2.1 Introduction

As mentioned in the previous chapter simple sequence repeats form an interesting group of genomic features owing to their property of length polymorphism arising out of insertions and deletions of their repeat units which occur mostly as a consequence of strand slippage during DNA replication. It is interesting, in the first place, to identify them from a DNA sequence in order to characterize their function. In the pre-genomic era, the technique followed used to be time consuming and cumbersome. This involved a number of steps such as, isolation of flanking regions specific for each microsatellite loci by construction and screening of different libraries, detection of microsatellites containing clones, designing of locus specific PCR primers and amplification of the respective regions from different sources of genomic DNA. However, in the post-genomic era availability of whole genome sequences has made the task of screening genomes for microsatellites simple and easy. A computer program specially developed for microsatellite screening is sufficient to scan the whole genome for sequence composition of motif, location and frequencies of microsatellite tracts.

As discussed in the previous chapter there are a few computational tools available for extraction of microsatellites from DNA sequences. All these tools focus on repeat extraction and do not automatically give location of repeats with respect to the coding/non-coding regions which forms an essential information as it can give an idea of possible role of microsatellite mutations on the coding regions nearby with respect to their transcription as well as translation. We therefore developed a program called SSRF which can extract enough information from genomes pertaining to microsatellites.

At the start of work a number of databases were available on the world wide web (WWW), which gave information on the DNA repeats used as genetic markers. Although the databases were very comprehensive, they were not infor-

mative enough to give insights towards understanding the role of microsatellites in pathogen's adaptability, virulence etc.. We therefore developed a relational database of microsatellites, called MICdb (Sreenu et al. 2003a) which forms, to the best of our knowledge, the first [1] comprehensive database of prokaryotic and viral microsatellites available of WWW. We also developed a web-server called MICAS (Sreenu et al. 2003b) to host the database and other information on microsatellites. This chapter gives the details of the program SSRF, the database MICdb and the web-server MICAS.

2.2 Simple sequence repeat finder (SSRF)

We have employed brute-force approach for finding perfect microsatellites. Since the algorithm is not dictionary based like TROLL (see previous chapter), this program does not require any apriori knowledge of repeat motif's sequence or repeat number. This method does not use pre-compiled list of motifs to check for repetitions. All the repeat sequences are generated on the fly without any limit to the repeat number.

2.2.1 The algorithm

The algorithm consists of two components

1. repeat detection

2. repeat annotation

[1] Fleche et al. developed a database of minisatellites in available genomes (http://minisatellites.u-psud.fr) (Fleche et al. 2001). This database was built using Tandem Repeat Finder program (Benson, 1999). This database organises minisatellites data as different genomic groups, like archaea, bacteria, eukaryota and virus.

Repeat detection

This makes use of the sliding window technique for repeat finding. To begin with a segment of the DNA sequence at the start of the 5′ end, of length S, (where S = 6 to begin with) is taken as a window and is assumed as the repeat unit which is checked for internal repeats to make sure that the repeat unit itself is not a repeat tract. For example, ATGATG. Here, the unit itself is a tract of ATG repeating twice. If the repeat unit does not contain internal repeats the next successive window of the same size is compared with the repeat unit. If the repeat unit (in other words, the first window) is same as the next successive one, the entire tract of the two repeats (windows) is considered as the microsatellite with repeat number two. Next successive window is also compared in a similar way (Figure 2.1). This process is continued until a non-identical window is encountered. On identifying a non-identical window, information on the repeat motif, its location and its repeat number are recorded. This scanning is carried out till the end of the sequence is encountered. This cycle thus constitutes the first frame scan for microsatellites with repeat units of size S. The second frame scanning starts by skipping one nucleotide at the start of the sequence at the 5′ end. Similarly scanning of all the frames are carried out by skipping one nucleotide at a time until the number of scanning cycles N is equal to S. The next cycle of scanning of genome, a window length equal to S-1 is considered and the scanning is carried out for all frames as mentioned above. The program is continued untill the searching cycles are completed for window of one nucleotide (see Figure 2.2 that depicts the program flow chart). During these cycles, care is taken to avoid redundancies in repeats. For this, any locus marked as a microsatellite in the previous scans is skipped in the subsequent scans.

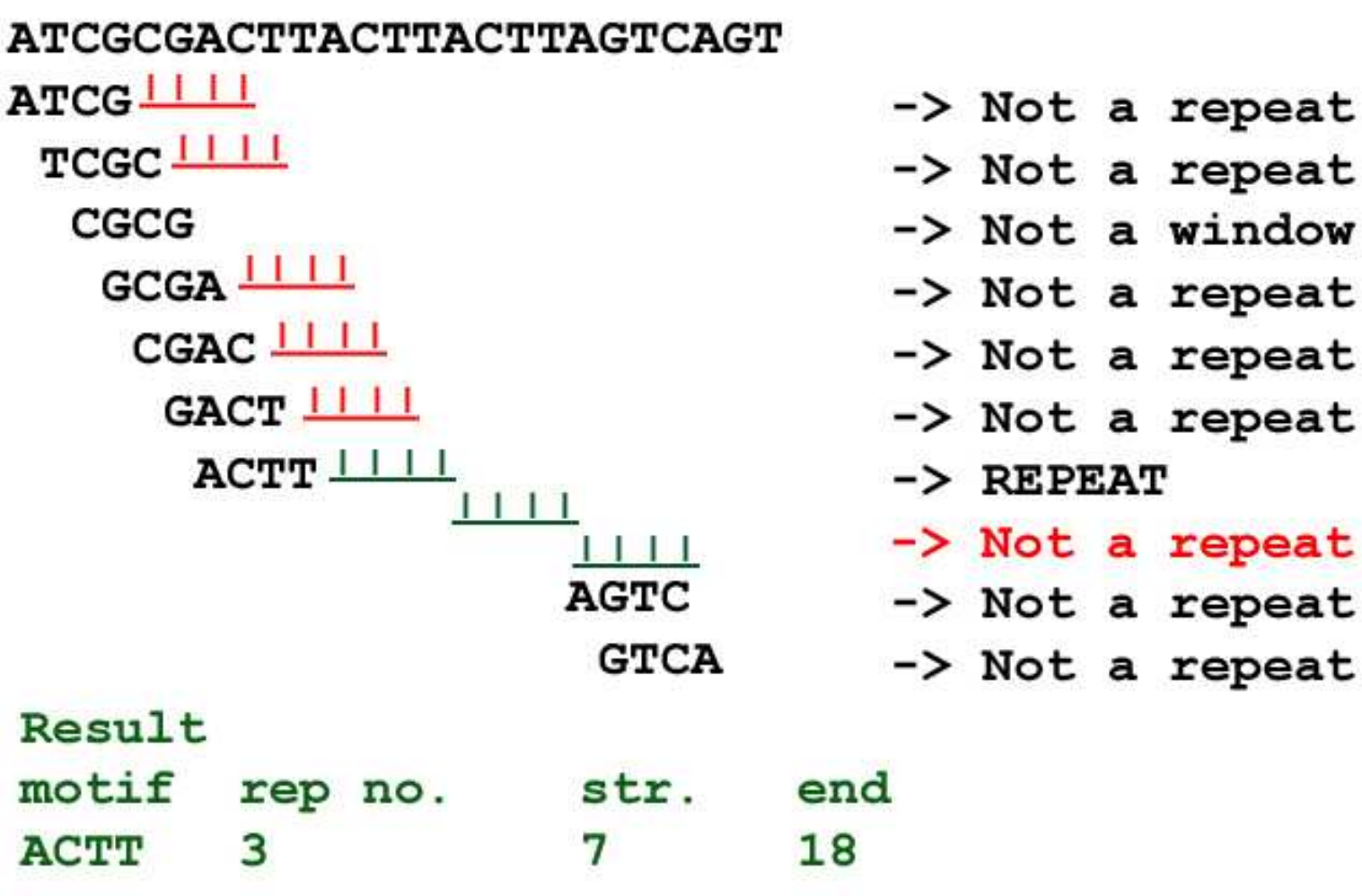

Figure 2.1: SSRF's Repeat finding method

Repeat annotation

All the microsatellites are recorded with locations with respect to the coding regions nearby. If the microsatellite is found in the non-coding region its precise location with respect to the upstream and downstream coding regions is recorded. SSRF uses the GenBank's ORF sequences file (.ffn) for obtaining the boundaries of the ORF. Repeats in the coding and non-coding regions are marked as 'C' and 'N' respectively. Any tract starting in the non-coding region and ending in the coding region or vice versa is marked as 'P' (partial) in the output.

2.2.2 Algorithm implementation

This algorithm has been implemented in SSRF using the C programing language. The C language was chosen owing to its fast computational power and its structured nature. The program takes the genome sequence file accession number as a command line argument and searches for the .fna file (sequence file) and the .ffn file (ORFs file) in the local directory. The program gives an error message "Genome sequence file not present" if the genome sequence file is not found and it gives a warning message "Coding region file not present" if the annotation file is not found. In the latter case, SSRF extracts the repeats, but the repeat location with respect to the coding and non-coding regions is not reported. A typical output of the SSRF program with and without the annotation file is shown in Fig. 2.3.

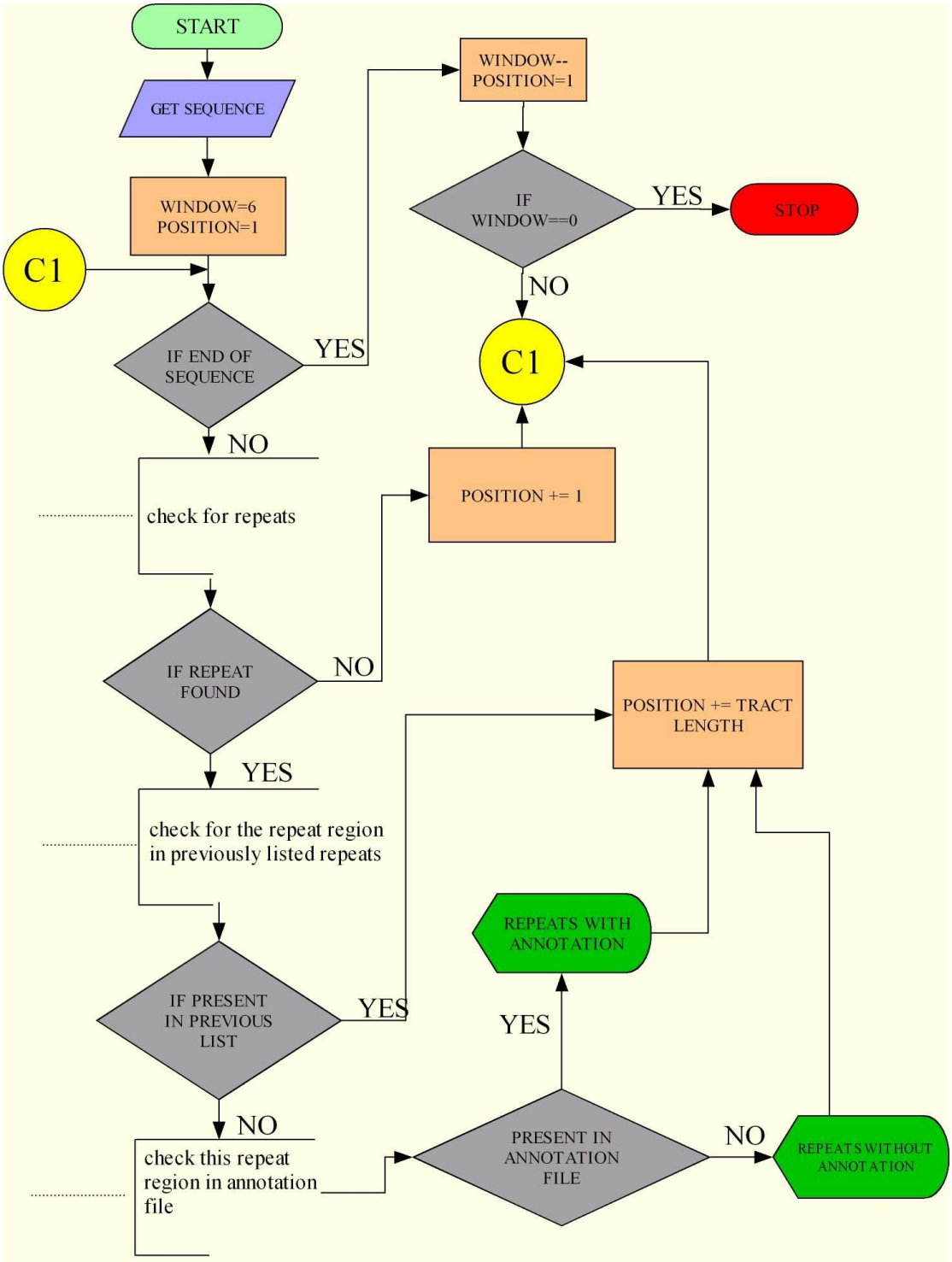

Figure 2.2: Flow-chart of Simple Sequence Repeat Finder(SSRF)

Input sequence

```
>gi|15607142| Mycobacterium tuberculosis H37Rv
TTGACCGATGACCCCGGTTCAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACTTAACGGCGACC
CTAAGGTTGACGACGGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTG
GCTCAATCTCGTCCAGCCATTGACCATCGTCGAGGGGTTTGCTCTGTTATCCGTGCCGAGCAGCTTTGTC
CAAAACGAAATCGAGCGCCATCTGCGGGCCCCGATTACCGACGCTCTCAGCCGCCGACTCGGACATCAGA
TCCAACTCGGGGTCCGCATCGCTCCGCCGGCGACCGACGAAGCCGACGACACTACCGTGCCGCCTTCCGA
```

Output (dinucleotide repeat) without annotation file (truncated output)

```
CA      2       30      33
GT      2       34      37
CG      2       42      45
TC      2       50      53
TC      2       105     108
GC      2       110     113
TC      2       147     150
CT      2       182     185
GC      2       225     228
```

with annotation file

```
CA      2       30      33      C
GT      2       34      37      C
CG      2       42      45      C
TC      2       50      53      C
TC      2       105     108     C
GC      2       110     113     C
TC      2       147     150     C
CT      2       182     185     C
GC      2       225     228     N
```

Figure 2.3: A typical output of the SSRF program with and without the annotation file

The first column in the result file is the repeating motif, the second column is the repeat number, and the third and fourth columns are the repeat start and end positions respectively. The fifth column gives information whether the tract is found in the coding region (C) or non-coding region (N).

The program has been tested on the RedHat Linux 8.0 and higher versions series machines. The typical execution time for extracting all the repeats (motif length from 1-6bp) from a bacterial genome is 5-10 sec. With the SSRF, it takes about 6 seconds to extract all the repeats from the *M. tuberculosis* H37Rv genome (4411529bp) on a Linux (RedHat 8.0) system with a Pentium 4 processor and 512MB RAM.

2.3 MICdb (Microsatellites Database) - The architecture

MICdb embodies a combination of relational and flat-file architectures. Precompiled non-redundant microsatellites obtained by means of whole genome scanning by SSRF, and the data pertaining to coding regions are stored in the relational database using MySQL4.0.14 (http://www.mysql.com). Data are stored in two different types of tables. Type-I table (Table 2.1) contains microsatellite motifs, repeat number, start and end positions and genomic location with respect to coding and non coding regions. Type-II table (Table 2.2) stores data pertaining to coding regions (protein ID, ORF start and end positions, ORF's functions, locus-tag IDs for database cross references and coding strand information). The data on each genome are stored in seven MySQL tables; one table (type-II) for coding regions information and six type-I tables for microsatellites mono to hexanucleotide respectively. Hence currently the database holds 3410 MySQL tables, 617594 flat-files and a master table containing all the genome codes with genome names and their group, ie. Archaea, Gram positive etc. The current composition of the database is given in the Table 2.3. The database is periodically updated using a JAVA program developed for this purpose.

Table 2.1: Description of type-I MySQL table that was used for storing microsatellite's information

Field	Type	Null	Key	Default	Extra
motif	varchar(15)	YES		NULL	
repeat	int(2)	YES		NULL	
sp	int(11)	YES		NULL	
ep	int(11)	YES		NULL	
region	char(1)	YES		NULL	

Table 2.2: Description of type-II MySQL table that was used for storing ORF's information

Field	Type	Null	Key	Default	Extra
PROT_ID	varchar(50)	YES		NULL	
PROT_DESC	varchar(255)	YES		NULL	
ORF_SPOS	int(11)	YES		NULL	
ORF_EPOS	int(11)	YES		NULL	
SWISS_ID	varchar(50)	YES		NULL	
STRAND	char(1)	YES		NULL	

Table 2.3: The current composition of MICdb

Number of Genomes		487
	Animal/Plant virus	175
	Phages	112
	Archaea	21
	Gram +ve bacteria	71
	Gram -ve bacteria	108
Number of records		150938723
	Animal/Plant virus	2777589
	Phages	1154557
	Archaea	11424277
	Gram +ve bacteria	48493946
	Gram -ve bacteria	87088354
Number of flat files		617594
	Sequence files	599
	Summary files	599
	ORFs+SST* files	616396

*SST = Secondary structure

In addition to MySQL tables, each genome has three flat-files. The first flat-file contains the complete genome sequence, the second file stores information of all the translated ORFs along with their secondary structures predicted using PSIPRED (McGuffin et al. 2000) and the third file contains the genome-wise summaries (genome composition, number of ORFs and coding density). Overall there are 1470 flat-files in the database. The complete relational schema of the database and data-flow therein are shown in Figure 2.4.

2.4 MICAS: Microsatellites Analysis Server

MICdb has been hosted on a web server called MICAS, which has been developed using JSDK2.0. The JAVA servlets of MICAS running on top of the apache web server (http://www.apache.org) handles queries and retrieves data from the database. The JAVA servlets have been designed to generate HTML pages dynamically, containing various microsatellite features, as responses to user queries.

MICAS server has a three-module architecture as illustrated in Figure 2.5. The first module is a processing unit (PU), the second module is MICdb and the third module is the user interface (UI). PU is made up of three programs W-SSRF, AUTOPRIMER and the processing engine.

2.4.1 W-SSRF: Web-based Simple Sequence Repeat Finder

Since our SSRF program that was developed in C progaming language was not compatable to host it on the web, we developed W-SSRF (Web-SSRF). The W-SSRF has been developed using Java programming language. The program scans a given nucleotide sequence for the presence of perfect simple sequence repeats. The extracted information pertaining to SSRs includes the sequence content of the motif, repeat numbers and start and end positions of the SSR tracts in the sequence.

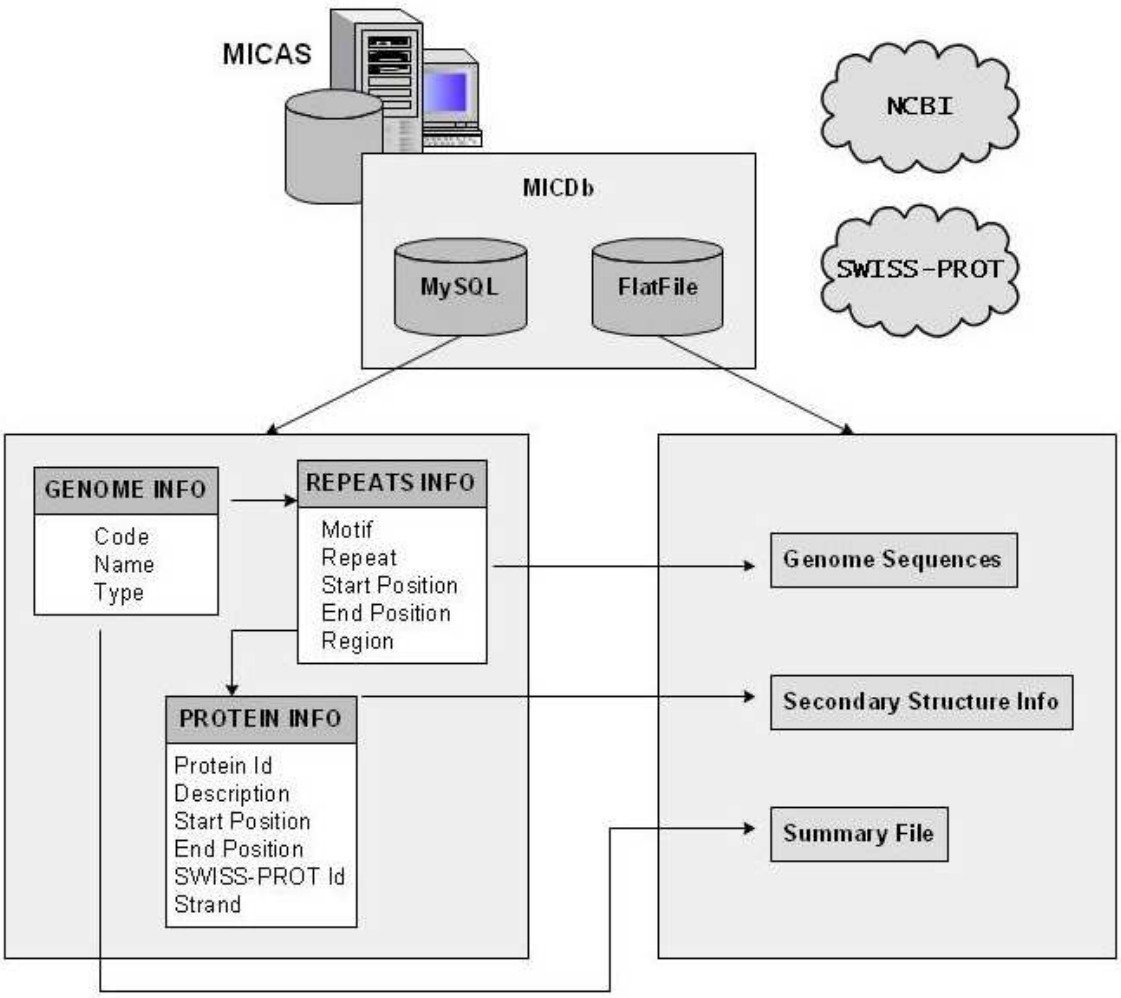

Figure 2.4: Entity relation (ER) diagram of MICdb

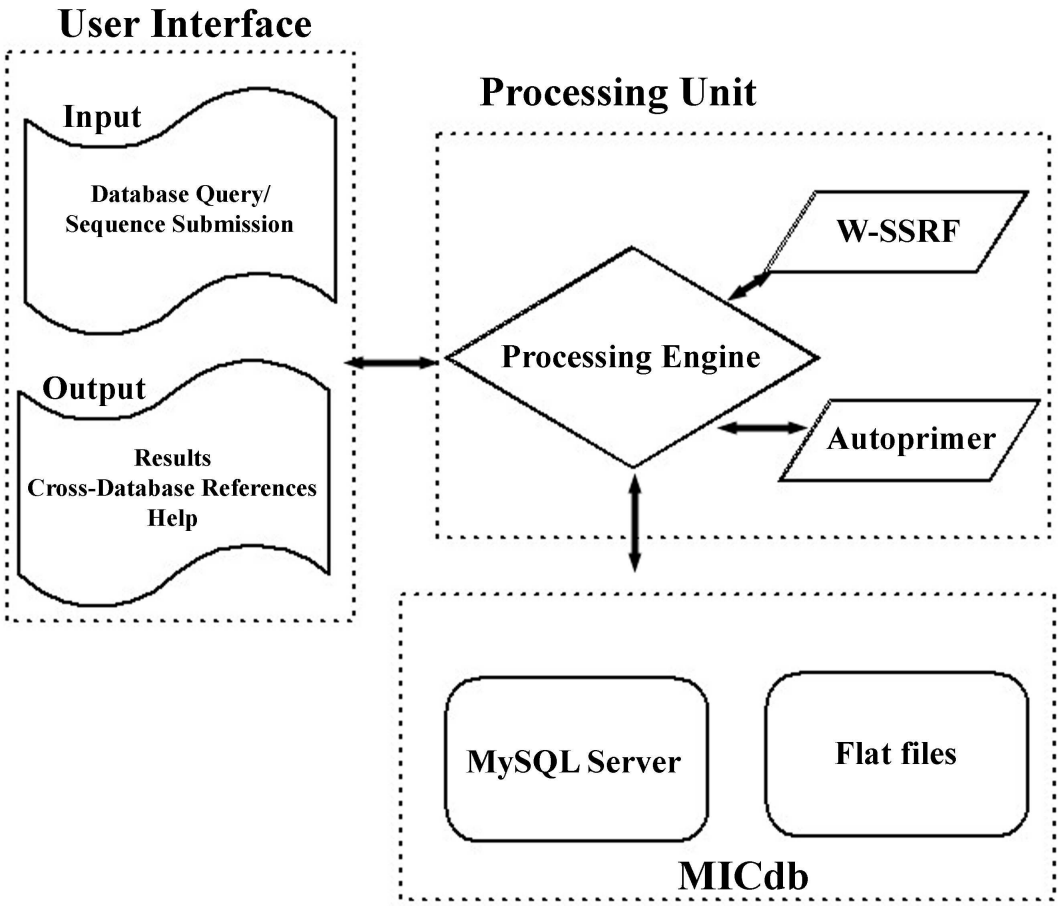

Figure 2.5: MICAS architecture

2.4.2 AUTOPRIMER: The primer design software

MICAS server also hosts a primer design tool called AUTOPRIMER to facilitate design of primers for a selected microsatellite loci. AUTOPRIMER has been developed in JAVA programming language. The essential parameters required for optimum design of primers viz., primer length, GC content and melting temperature (Tm) have been considered. Tm is calculated using nearest neighbour thermodynamics equation (Breslauer et al. 1986). Default values for primer length and GC percentage in the primer sequences have been set to 20 nucleotides (optimum range is 18-30) and 50 (optimum range is 45-65), respectively. Special care has been taken to avoid palindromes and repetitive sequences in the primers, complementarity within the forward and reverse primers and also to check for their dimerisation potential (Kampke et al. 2001). Each primer is further tested for unintended hybridisation with itself, by testing for self-annealing and self-end annealing. Complementarity between primers is indicated by score value. High score is given to highly complementary primer pairs (Kampke et al. 2001). By default, the program generates 25 sets of possible primer pairs for a given sequence.

The processing engine of PU processes MICAS output and AUTOPRIMER's output according to users choice of microsatellites having specified repeat size and number.

2.4.3 User interface

The third module is the user interface, developed using Java servlets. The interface provides a platform for MICdb, W-SSRF and the Autoprimer to input queries as well as to display query results by means of dynamically generated HTML pages. To query the MySQL database, the MM.MySQL JDBC driver has been used.

2.5 Data Retrieval from MICdb Using MICAS

There are three modes of query facilities:

1. Classic query
2. ORF-wise query
3. Advanced quey

Classic query facility

MICAS has a very user friendly and interactive front-end using which MICdb can be queried. User can select a genome from the drop-down menu and query the database for occurrence of tandem repeats of microsatellite motifs of specified size (S), at least specified number of times (N). Following the query MICAS outputs a table containing the complete list of microsatellite motifs satisfying S and N. This table also gives against each motif the number of times it occurs in the whole genome and the minimum and maximum number of times it is tandemly repeated at various loci. The motifs in this table are hyper-linked texts by clicking which further details of the motifs can be obtained. The details of the motifs are, their genomic co-ordinates (starting and ending nucleotide numbers) and regions of their occurrence (whether they are in coding or non-coding regions). If the microsatellite tracts occur in the coding region, the details of that region are also provided.

ORF-wise query

This is a simple query facility where the database can be searched for ORFs containing specific microsatellites by specifying either gene-id or gene name. Query will be added with the prefix and suffix wildcard characters by default. This not only makes the users job easier even if they do not know the complete gene-id

number or gene name, but also helps in avoiding the problems of typos, by giving short keywords. The initial output page is a summary page which shows the number of hits (i.e., number of genomes) containing the ORFs with matching gene-id or gene name. User can further navigate into the genomes and extract microsatellite information.

Advanced query

This is the query through which a user can explore maximum of the database. Choosing a particular motif of interest, user can make a complex query by clubbing multiple select options like single repeat number or all repeats, region of choice (C coding, N- non-coding or both), and genome of interest. Selection of genome is provided with a wide range of choices as a user can select only one genome or two/more or all genomes from the three types of genome groups (Archaea, Eubacteria and Animal/Plant virus) available. The initial query result page carries a summary of the search results, so that user can make a step by step walk through the results.

The microsatellite features that are generated upon user queries are: (i) Microsatellite tracts annotated with information corresponding to coding/non-coding/partially coding regions, (ii) Graphical illustration of relative position of microsatellite tracts occurring in non-coding regions flanked by upstream and downstream coding regions and (iii) Illustration of the coding regions containing microsatellites, annotated with secondary structural information. Furthermore, every page that is dynamically generated is also provided with sufficient textual information to help data analysis and further user navigation. The coding regions are hyperlinked to their respective entries in GenBank and SWISSPROT in order to provide annotations available in those databases.

If user wishes, he/she can select any of the microsatellite tracts (obtained from database query or sequence scanning) along with its flanking sequence for AUTOPRIMER to design primers for PCR. Provision has been made for the user

to change the default settings of various parameters for primer design. By default AUTOPRIMER generates 25 sets of primer pairs.

2.5.1 Illustration with an example query

A sample query flow using classic query facility has been demonstrated in the Figure 2.6. In this example, gram positive bacterium *B. sabtilis* is queried for tetra nucleotide microsatellite with a minimum repeat number 3. Query has returned 62 microsatellite tracts presented in a table form. Among these tracts, GGCA motif is selected and clicked for further details. The details returned from the database can be summerised as follows: (a) there are only two microsatellites with GGCA sequence motif, (b) one of the microsatellites occurs in the coding region and the other in the non-coding region. Subsequent query to the database yields information that the coding region is annotated as a primosomal replication factor Y and the repeat tract is present in the junction of a helix and a loop. The second microsatellite occurs in the non-coding region which is in between the coding regions of extracellular serine protease and negative regulatory protein of SacY. Subsequent query results in a graphical illustration showing the relative position of the microsatellite with respect to the upstream and downstream coding regions. Any of these microsatellite tracts can be chosen for primer design. The figure also illustrates usage of AUTOPRIMER for design of primers.

Since its launching on the WWW, MICdb has received a number of hits and also has been cited by some publications and the details are given in Table 2.4 and Table 2.5.

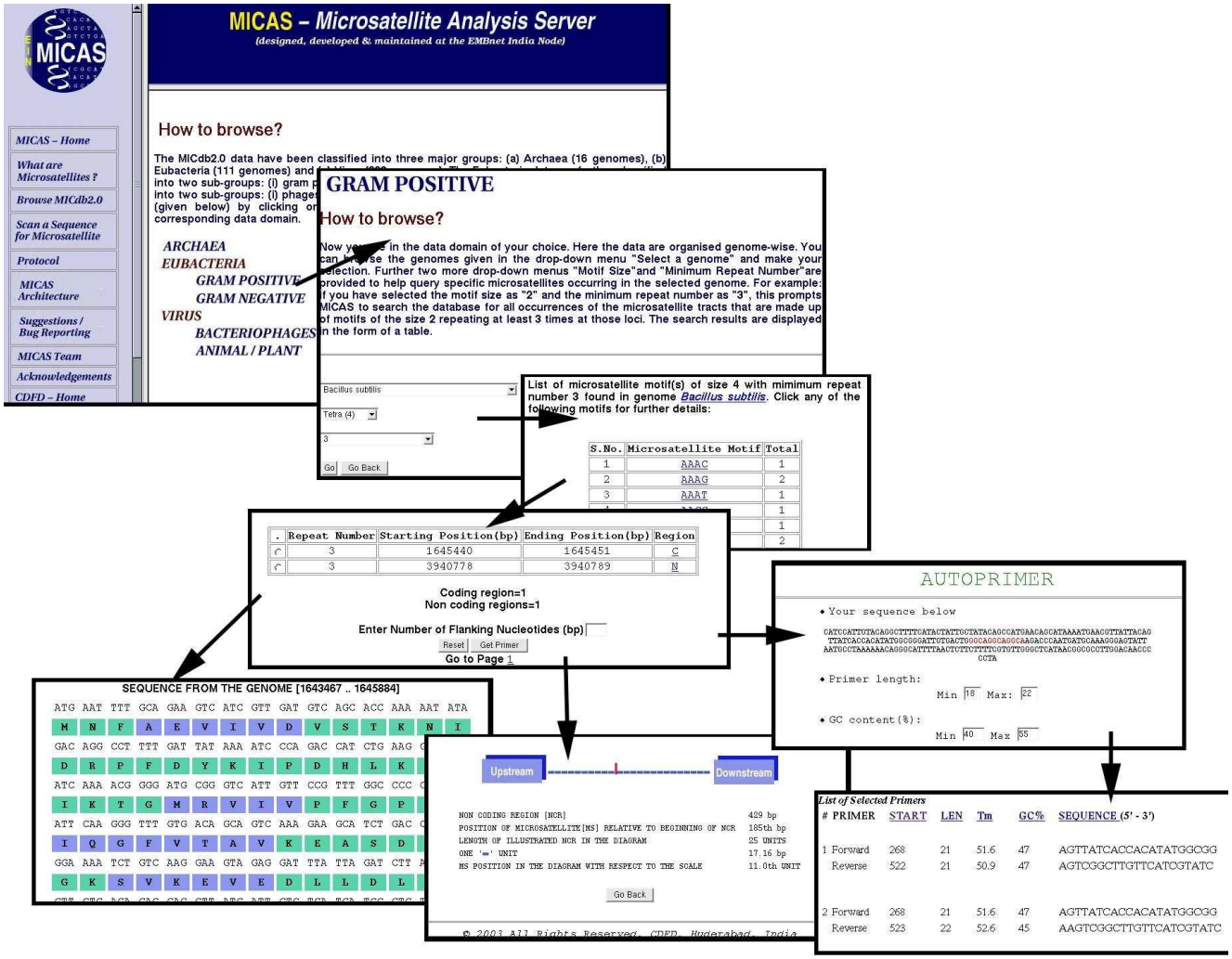

Figure 2.6: Illustration of information extraction from MICdb using classic query facilisty

Table 2.4: MICdb access statistics for the last six months

Month	Number of hits
June 2005	588
July 2005	609
August 2005	554
September 2005	388
October 2005	532
November 2005	626

Table 2.5: Citations for MICdb

McAdams et al. (2004) *Nat. Rev. Genet.* **5:** 169-178
Mizuta et al. (2004) *Chem-Bio Inform. J.* **4:** 133-141
Groathouse et al. (2004) *J. Clinical Microbiol.* **42:** 1666-1672
Boby et al. (2005) *Bioinformatics* **21:** 811-816
Prasad et al (2005) *Nucleic Acids Res.* **33:** D403-D406

2.6 Summary

In this chapter we gave details of the SSRF program, MICdb and MICAS. MICdb forms the first relational database of prokaryotic and viral microsatellites on the WWW. As microsatellites have been considered as critical elements in many of the biological events, the ready availability of microsatellites and their information related to coding and non-coding region makes this database very unique. Strengths of MICdb are cross-database references on microsatellite containing coding regions and coupling the database with primer design software AUTOPRIMER and graphical illustration of information pertaining to occurrences of microsatellite tracts in coding and non-coding regions. These together give a lot of insight into possible functional roles of the microsatellites in the genomes of interest.

Chapter 3

Analysis of mycobacterial genomes for distribution and abundance of microsatellites

---

Publications from this chapter

- **Sreenu, V.B.**, Nagaraju, J. and Nagarajaram, H. A. (2005) Survey and analysis of microsatellites in mycobacterial genomes (communicated).

3.1 Introduction

Microsatellites owing to their property of length polymorphism render the microbes some novel functionalities related to adaptation, virulence, phase variation, multi-drug resistance etc. As mentioned in the previous chapters these sequence features are present in all the genomes known so far and several studies have been reported including some excellent reviews which describe distribution of microsatellites in several prokaryotic and eukaryotic genomes (Moxon et al. 1994; Usdin and Grabczyk, 2000; van Belkum et al. 2000;). It is now established that microsatellite loci are involved, directly or indirectly, in the pathogenicity of the pathogens such as *Haemophilus influenzae* and *Nesseria gonorrhoeae* (Moxon et al. 1994; van Belkum et al. 1998). However, a look at the literature gives a revelation that the pathogens viz., mycobacteria have not been analyzed for abundance and distribution of microsatellites, although there are some reports on insertion elements and MIRU that are presently in use as genetic markers (Cole et al. 1998; Hermans et al. 1991; Kamebeek et al. 1997; van Soolingen et al. 1993). In the light of this it seemed important to analyse the mycobacterial genomes for distribution, frequencies and polymorphisms of microsatellites.

Currently, complete genome sequences for the five mycobacteria namely, *M. avium* (Li et al. 2004), *M. leprae* (Cole et al. 2001), *M. bovis* (Garnier et al. 2003) and two strains of *M. tuberculosis* (CDC1551 (Fleischmann et al. 2002) and H37Rv (Cole et al. 1998)) are available in the public domain. *M. avium* is a common bacterium in surface water and soils and is the causative agent of the Crohn's disease in humans (Cosma et al. 2003). *M. leprae* causes leprosy in humans, while *M. bovis* is the causative agent of tuberculosis in many animals and humans. *M. tuberculosis* is the major cause of tuberculosis in humans. All these organisms are GC-rich genomes. Approximate coding density of these genomes is about 90%, except in *M. leprae*, where it is 49% (Cole et al. 2001) (Table 3.1).

Table 3.1: Mycobacterial genomes that are considered for microsatellites analysis

Orgnism	**Genome Size (bp)**	**GC (%)**	**Coding density (%)**	**Reference**
M. avium	4829781	69	91	Li et al. (2004)
M. bovis	4345492	66	90	Garnier et al. (2003)
M. leprae	3268203	58	49	Cole et al. (2001)
M. tuberculosis CDC1551	4403836	66	90	Fleischmann et al. (2002)
M. tuberculosis H37Rv	4411529	66	90	Cole et al. (1998)

As mentioned earlier, the mycobacterial genomes lack the post-replication DNA repair enzymes *mutL, mutS* and *mutH* (Mizrahi and Andersen, 1998; Springer et al. 2004) and therefore it can be surmised, as a null hypothesis, that the mutations in microsatellites occur as unregulated events and therefore the genomes are enriched with long microsatellites. In order to test this null hypothesis we analysed the five mycobacterial genomes for frequencies and distributions of microsatellites and the results are reported in this chapter.

3.2 Methods

We considered a sequence tract as a microsatellite where a motif of size between 1-6 bp tandemly repeats itself at least two times at a locus. All the microsatellites considered in this study unless otherwise mentioned, are the perfect microsatellites.

The microsatellite data pertaining to the five mycobacterial genomes *M. avium*, *M. bovis*, *M. leprae*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv comprising of sequence of the repeat motif, repeat size, repeat number and location with respect to the coding and non-coding regions were extracted from the corresponding genome sequences using SSRF. The observed number of each class of microsatellites (mono, di, etc.) in a genome was compared to the number that could be expected by chance in a randomized genome of the same length and composition. Each of the genome sequence was randomized ten times, and from the ten randomized samples the average numbers of microsatellites were calculated as the expected numbers. We used SHUFFLESEQ program available in the EMBOSS software suite (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) for randomization of the genomes. Statistical significance of the observed number of microsatellites as compared to the expected number was carried out with students t-test (programs were written in C with functions taken from Numerical recipes in C (Press et al. 1992)).

3.3 Results

Every genome, except *M. leprae*, harbors as many as one million microsatellites. The *M. leprae* genome harbors 25% less as compared to the other genomes, however, the net genomic occupancy of the microsatellites expressed as the ratio of the number of bases in the microsatellites and in the whole genome, is similar to that found in the other genomes (Table 3.2).

3.3.1 Microsatellite density profile

Figure 3.1 shows a plot of the number of microsatellites per ten thousand bases (referred to as tract density) found in each of the genomes. The tract densities vary in the range of 2200-2300 tracts/10Kb. A look at the tract density profiles reveals that the microsatellites are scattered throughout the mycobacterial genomes. The profiles of the genomes other than *M. leprae* comprise of regions conspicuously either rich or poor in microsatellites; the *M. leprae* genome is nearly devoid of such regions. The profiles of *M. avium* and *M. leprae* stand distinctly different from *M. tuberculosis* CDC1551, while those of *M. tuberculosis* H37Rv and *M. bovis* are very similar, indicating homology of microsatellite evolution. *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv as well as *M. bovis* genomes harbor more number of microsatellite rich regions than microsatellite poor regions; *M. avium* harbors more number of microsatellite poor regions than microsatellite rich regions. Such regions are missing in *M. leprae*.

An examination of the microsatellite rich and poor regions in the five genomes revealed that a large majority of ORFs (32 out of 37) in the microsatellite poor region of *M. avium* encode hypothetical proteins. In *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv and *M. bovis* genomes the microsatellite rich regions encode proteins belonging to the PE and PPE family of proteins in addition to some hypothetical and heat shock proteins (Table 3.3).

Table 3.2: The net genomic occupancy of the microsatellites expressed as the ratio of the number of bases in the microsatellites and in the whole genomes of mycobacteria

Genome	Total number of microsatellites in bp (a)	Genome size in bp (b)	Genomic occupancy (a/b)
M. avium	3476428	4829781	0.72
M. leprae	2118239	3268203	0.65
M. bovis	2994297	4345492	0.69
M. tuberculosis CDC1551	3032807	4403836	0.69
M. tuberculosis H37Rv	3038137	4411529	0.69

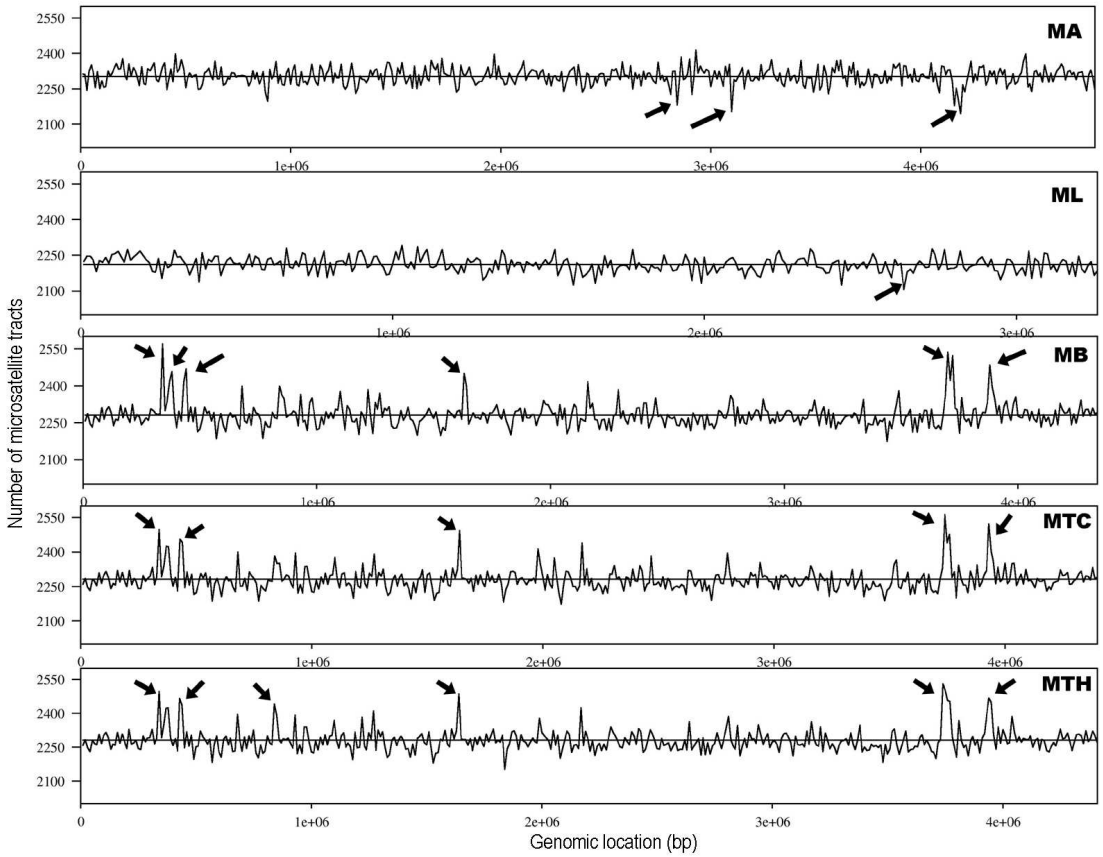

Figure 3.1: Tract density profile of the microsatellites in *M. avium* (MA), *M. leprae* (ML), *M. bovis* (MB), *M. tuberculosis* CDC1551 (MTC) and *M. tuberculosis* H37Rv (MTH). Tract density is equal to the number of tracts per 10Kb. Peaks higher or lower than three standard deviations from the mean are shown by arrows and referred to as conspicuously rich or poor tract densities respectively.

Table 3.3: The regions with conspicuously high/low tract densities and the ORFs found in those regions

Genome region (x 10kb)	Local GC %	Repeat rich (+)/Repeat poor (-)	Total orfs	Number of hypothetical proteins	Non-hypothetical proteins (GenBank ID)
M. avium					
283-284	67	-	9	7	ATP-dependent DNA/RNA helicase (dead) (41408619), lopoprotein (*lpre*) (41408620)
309-310	64	-	12	12	
415-416	63	-	9	9	
418-419	62	-	7	4	IS1110 (41409846), small membrane protein (MMPS1) (41409848), large membrane protein (MMPL4_5) (41409849)
M. leprae					
263-264	56	-	3	1	aminopeptidase I (*pepc*) (15828189), phosphoribosylformylglycinamidine synthase I (*purq*) (15828190)
M. bovis					
33-34	75	+	5	2	PE_PGRS3a(31791456), PE_PGRS3 (31791457), PE_PGRS4 (31791458)
37-38	62	+	6	2	PPE6 (31791484), oxidoreductase (31791485), conserved integral membrane protein(31791487), conserved exported protein (31791488)
43-44	64	+	6	3	adenylosuccinate synthase (*purA*) (31791534), conserved integral membrane protein (31791536), conserved membrane protein (31791538)
162-163	70	+	6	0	6-phosphogluconolactonase *devB* (*6pgl*) (31792639), putative oxpp cycle protein (*opcA*) (31792640), glucose-6-phosphate 1-dehydrogenase (*g6pD*) (31792641), transaldolase (31792642), transketolase (31792643), PE_PGRS27 (31792644)
369-370	72	+	2	1	PE_PGRS50a (31794528)

(continued)

Table 3.3: (continued...)

Genome region (x 10kb)	Local GC %	Repeat rich (+)/Repeat poor (-)	Total orfs	Number of hypothetical proteins	Non-hypothetical proteins (GenBank ID)
371-372	64	+	1	0	PPE56a (31794533)
387-388	76	+	3	0	fatty-acid-coa synthetase (31794682), PE_PGRS53 (31794683), PE_PGRS54 (31794684)
M. tuberculosis CDC1551					
33-34	72	+	11	9	PE_PGRS (15839660), PPE(15839665)
42-43	63	+	5	0	heat shock protein (*grpE*) (15839737), heat shock protein (*dnaJ*) (15839738), transcriptional regulator *hspR*(15839739), PPE (15839740), PPE (15839741)
43-44	64	+	7	5	adenylosuccinate synthetase (15839743), divalent cation transporter (15839748)
163-164	73	+	6	2	PE_PGRS (15840909), cytochrome c oxidase folding protein (15840911), PE_PGRS (15840912), quinone oxidoreductase (15840914)
216-217	60	+	5	1	PPE (15841389), PPE (15841389) acyltransferase family protein, lipoprotein (15841392)
373-374	71	+	4	3	PE_PGRS (15842940)
374-375	64	+	3	1	IS1608', transposase (15842945), IS1561', transposase (15842946)
375-376	64	+	1	1	
392-393	78	+	2	0	PE_PGRS (15843119), PE_PGRS (15843120)
M. tuberculosis H37Rv					
33-34	72	+	7	3	Acyl-coa synthase (*fadd27*) (15607416), PE_PGRS (15607419), PE_PGRS (15607420), PPE (15607421)

(continued)

Table 3.3: (continued...)

Genome region (x 10kb)	Local GC %	Repeat rich (+)/Repeat poor (-)	Total orfs	Number of hypothetical proteins	Non-hypothetical proteins (GenBank ID)
42-43	63	+	5	0	heat shock protein (*grpE*) (15607492), heat shock protein (*dnaJ*) (15607493), heat shock protein (*hspR*) (15607494), PPE (15607495), PPE (15607496)
43-44	64	+	7	5	adenylosuccinate synthase (*purA*) (15607498), magnesium ion transporter (*mgtE*) (15607503)
83-84	71	+	10	7	PE_PGRS (15607882), PE_PGRS (15607886), PE_PGRS (15607887)
163-164	73	+	5	1	PE_PGRS (15608588), cytochrome c oxidase assembly factor (*ctaB*) (15608589), PE_PGRS (15608590), quinone oxidoreductase (*qor*) (15608592)
373-374	67	+	2	0	PE_PGRS (15610480), PE_PGRS (15610481)
374-375	67	+	2	1	PPE (15610483)
375-376	65	+	3	2	PPE (15610486)
376-377	65	+	6	5	methylenetetrahydrofolate dehydrogenase (fold) (15610492)
393-394	75	+	4	1	PE_PGRS (15610644), Probable acetohydroxyacid synthase I large subunit (*ilvX*) (15610645), PE_PGRS(15610647)
394-395	79	+	3	0	PE_PGRS (15610648), Acyl-coa synthase (fadd18) (15610649), PE_PGRS (15610650)

3.3.2 Microsatellite distribution and abundance

The numbers of microsatellites of different repeat sizes (mono to hexa) found in each genome, are given in Tables 3.4 to 3.10 and the abundance of microsatellites in terms of the sequence motifs is given in Table 3.11.

From Table 3.4 it can be seen that the mononucleotide tract lengths in *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551 genomes seem to be highly restricted as compared to the other genomes, with repeat numbers never exceeding 9, while in the other three genomes, repeat numbers go up to 27 at some loci. Except for this difference all the five genomes show abundance of short tracts with repeat iteration of two that occur significantly more than expected, while tracts with repeat iterations greater than two are underrepresented. The ratio of the observed to the expected (O/E ratio) numbers of mononucleotide tracts decreases very steeply as the number of repeats in the tracts increases, indicating a strong selection against long mononucleotide tracts. It may also be noted that while the enriched short tracts of repeat number two occur in both coding as well as non-coding regions, the longer tracts occur only in the non-coding regions. About three-fourth of the mono tracts are G/C - a reflection of the GC richness of these genomes, and these are underrepresented (see Table 3.11).

The repeat numbers of dinucleotide microsatellites are restricted to a maximum of five or six in *M. avium*, *M. bovis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv whereas in *M. leprae* the repeat number at some loci goes up to 18 (Table 3.5). Similar to observations in mono tracts, the O/E ratio of di tracts in all five genomes falls with increase in the number of repeats. Tracts with repeat numbers less than six occur without any striking bias towards coding or non-coding regions. The long tracts found in *M. leprae* are confined to the non-coding regions. In all the genomes GC motifs occur more frequently, and the number is overrepresented (see Table 3.11). The least frequently found AT/TA motifs are underrepresented. For GC/CG, GA/AG, CA/AC and GT/TG repeat tracts, the *M. leprae* genome shows a distinct overrepresentation as compared to

the other genomes where they are underrepresented.

All the five genomes show uniform overrepresentation of trinucleotide repeat tracts, which hardly ever exceed five iterations of repeats per locus (Table 3.6). In fact, over representation increases (i.e. O/E ratio) as the microsatellite repeat number increases, indicating a strong selection for the accumulation of long trinucleotide tracts. It is interesting to note that trinucleotide microsatellites with repeat number greater than four are found in the coding regions in *M. avium*, *M. bovis* and *M. tuberculosis* H37Rv, while in *M. leprae* these repeats occur in non-coding regions. Most of the trinucleotide repeat motifs except GAG, TAA and TAG are overrepresented in all the genomes (see Table 3.11).

Of the higher order microsatellites (with repeat unit size between 4 and 6) only hexanucleotide tracts of all repeat numbers are consistently over represented in all the mycobacterial genomes (Tables 3.7 to 3.9). It is also interesting to observe that *M. leprae* shows a distinct overrepresentation of pentanucleotide tracts. In mycobacterial genomes, in general, there is a universal overrepresentation of GC rich motifs as compared to AT rich motifs (supplementary data).

From Tables 3.11 where the summary of microsatellites is given, it is clear that microsatellites with small repeat sizes are more abundant than those with large size repeats in all the mycobacterial genomes. The number decreases as the microsatellite repeat size increases; with mono tracts being the most abundant ones possessing tract densities in the range of 160-170 tracts per Kbp, and hexa tracts being the least abundant with a mere one repeat or less per Kbp. Although the genomic distribution of microsatellites is grossly uniform, the occurrence of tri and hexa repeats is slightly more biased towards the coding regions, while the occurrence of mono, di, tetra and penta is slightly more biased towards the non-coding regions.

Table 3.4: Number of mononucleotide repeat tracts observed in the mycobacterial genomes. The observed number of repeats is compared against the average number of repeats occurring in ten randomized samples of the same genome, and if the observed number is significantly more ($p<0.001$) than the expected number, it is overrepresented (+) and similarly if the observed number is significantly less than the expected number, it is underrepresented (-).

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	623349 +	91	560815 (548)
3	147016 -	89	183370 (407)
4	31201 -	87	61596 (233)
5	6264 -	85	20944 (101)
6	662 -	77	7267 (77)
7	37 -	56	2512 (53)
8	7 -	100	860 (31)
10	1 -	100	104 (14)
M. leprae			
2	437169 +	50	409758 (896)
3	93349	49	113941 (248)
4	20933	45	30822 (157)
5	4848	41	8330 (55)
6	850	33	2307 (19)
7	151 -	35	628 (17)
8	28	42	174 (13)
9	3	0	49 (5)
10	2	0	14 (4)
11	1 -	0	4 (2)
12	1	100	1 (1)
16	1 +	0	0
20	1 +	0	0
22	1 +	0	0
M. bovis			
2	579496 +	90	521051 (697)
3	130389	88	159780 (269)
4	27674 -	87	49585 (177)
5	5733	86	15761 (77)
6	813	85	5085 (65)
7	127	88	1641 (41)
8	8	75	533 (11)
9	1	100	175 (10)
11	2	100	18 (4)
15	1 +	100	0

(continued)

Table 3.4: (continued...)

27	1 +	0	0
Repeat number	Observed	% in coding	Expected (sd)
M. tuberculosis CDC1551			
2	587333 +	90	528649 (404)
3	132130	88	161935 (312)
4	28023	87	50234 (204)
5	5786	85	15949 (100)
6	826	85	5098 (75)
7	138	81	1658 (41)
8	6	66	527 (10)
9	1	100	179 (14)
M. tuberculosis H37Rv			
2	588352 +	90	529032 (460)
3	132525	89	162389 (382)
4	28137	87	50343 (179)
5	5774	86	15860 (112)
6	824	86	5174 (72)
7	137	86	1654 (41)
8	5	100	531 (18)
9	2	50	170 (16)

Table 3.5: Number of dinucleotide repeat tracts observed in the mycobacterial genomes

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	152127 -	91	154813 (282)
3	10281	90	13286 (131)
4	1043	90	1316 (35)
5	69	92	146 (12)
M. leprae			
2	94326 -	49	103624 (301)
3	4825	46	6819 (65)
4	260	52	464 (28)
5	29	41	35 (8)
6	4 +	25	2 (1)
7	2 +	0	0
8	3 +	0	0
9	1 +	0	0
10	2 +	0	0
11	1 +	0	0
14	1 +	0	0
15	1 +	0	0
17	1 +	0	0
18	1 +	0	0
M. bovis			
2	129978	89	138647 (303)
3	7580	88	10671 (118)
4	599	90	932 (18)
5	43	90	80 (8)
6	1	100	9 (3)
M. tuberculosis CDC1551			
2	131690	89	140518 (232)
3	7691	88	10796 (90)
4	607	90	922 (20)
5	45	88	92 (14)
M. tuberculosis H37Rv			
2	131866	89	140732 (314)
3	7710	88	10829 (100)
4	605	90	934 (30)
5	45	88	84 (6)

Table 3.6: Number of trinucleotide repeat tracts observed in the mycobacterial genomes

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	106819 +	93	63768 (211)
3	6592 +	94	1807 (40)
4	420 +	96	61 (3)
5	15 +	100	3 (1)
7	1 +	100	0
M. leprae			
2	50458 +	51	35925 (232)
3	1649 +	57	628 (25)
4	39 +	76	10 (2)
5	6 +	0	0
6	1 +	0	0
9	1 +	0	0
21	1 +	0	0
M. bovis			
2	84224 +	91	53224 (117)
3	4066 +	93	1264 (43)
4	251 +	95	38 (5)
5	28 +	100	1 (1)
M. tuberculosis CDC1551			
2	85166 +	91	53873 (150)
3	4117 +	92	1269 (42)
4	244 +	91	36 (6)
5	37 +	81	1 (1)
6	1 +	100	0
M. tuberculosis H37Rv			
2	85287 +	92	54098 (142)
3	4103 +	94	1290 (29)
4	243 +	96	38 (6)
5	32 +	100	1 (0)
6	1 +	100	0
7	1 +	100	0

Table 3.7: Number of tetranucleotide repeat tracts observed in the mycobacterial genomes

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	15871 -	89	19265 (112)
3	119	89	173 (8)
4	7 +	42	2 (1)
M. leprae			
2	9421	47	9443 (112)
3	32	50	42 (6)
M. bovis			
2	13165	89	15266 (135)
3	68	85	107 (10)
4	1	100	1 (1)
M. tuberculosis CDC1551			
2	13294	89	15450 (96)
3	70	88	109 (9)
4	1	100	1 (1)
M. tuberculosis H37Rv			
2	13312	89	15569 (108)
3	68	86	110 (10)
4	1	100	1 (1)

Table 3.8: Number of pentanucleotide repeat tracts observed in the mycobacterial genomes

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	6102 -	89	6250 (92)
3	18	77	22 (7)
M. leprae			
2	3030 +	47	2632 (56)
3	8 +	50	2 (2)
M. bovis			
2	4347	88	4651 (89)
3	9	88	9 (2)
M. tuberculosis CDC1551			
2	4424	88	4687 (48)
3	10	80	10 (3)
M. tuberculosis H37Rv			
2	4457	88	4756 (79)
3	9	88	9 (2)

Table 3.9: Number of hexanucleotide repeat tracts observed in the mycobacterial genomes

Repeat number	Observed	% in coding	Expected (sd)
M. avium			
2	4438 +	94	1724 (46)
3	24 +	100	1 (1)
M. leprae			
2	1323 +	52	676 (20)
3	4 +	100	0
7	1	100	0
M. bovis			
2	2727 +	93	1259 (28)
3	18 +	100	0
4	3 +	100	0
M. tuberculosis CDC1551			
2	2750 +	90	1260 (32)
3	19 +	94	0
4	2 +	100	0
M. tuberculosis H37Rv			
2	2748 +	94	1289 (25)
3	18 +	94	1 (1)
4	2 +	100	0

Table 3.10: Number of microsatellites found in the mycobacterial genomes (a summary)

Motif	*M. avium*	*M. leprae*	*M. bovis*	*M. tb* CDC1551	*M. tb* H37Rv
MONO	808538 - (837824)	557338 - (566029)	744245 - (753695)	754243 - (764316)	755756 - (765236)
DI	163520 - (169582)	99457 - (110944)	138201 - (150339)	140033 - (152337)	140226 - (152587)
TRI	113847 + (65639)	52155 + (36563)	88569 + (54527)	89565 + (55180)	89667 + (55427)
TETRA	15997 - (19440)	9453 (9486)	13234 - (15374)	13365 - (15559)	13381 - (15680)
PENTA	6120 - (6272)	3038 + (2635)	4356 - (4660)	4434 - (4697)	4466 - (4765)
HEXA	4462 + (1725)	1328 + (676)	2748 + (1260)	2771 + (1260)	2768 + (1290)

Table 3.11: Number of occurrences of individual repeat motifs in mycobacteria (di and trinucleotide repeat motifs are listed as a group i.e TG = GT/TG). Occurrence of these repeats is compared against the repeats from ten random genomes and significant over-representation and under-representation are indicated as + and - symbols respectively. Numbers in parenthesis denote average number of repeat tracts in the ten random genomes.

	M. avium	*M. leprae*	*M. bovis*	*M. tb CDC1551*	*M. tb H37Rv*
Mononucleotide motifs					
A	93081 + (92777)	108332 + (107064)	105058 + (100988)	106686 + (102608)	106821 + (102915)
T	91709 (91870)	109980 + (108417)	105645 + (101191)	107067 + (102771)	107129 + (102782)
G	311170 (326036)	171119 (176817)	264439 (274941)	267766 (278569)	268608 (279150)
C	312578 (327141)	167907 (173731)	269103 (276575)	272721 (280368)	273198 (280389)
Dinucleotide motifs					
AT	1637 (4346)	7707 - (9814)	4164 - (6094)	4229 - (6175)	4227 - (6108)
GC	101590 + (86677)	32346 + (31067)	71466 + (64605)	72362 + (65480)	72409 + (65720)
AG	11919 (19684)	10928 - (17552)	11617 - (19876)	11789 - (20123)	11853 - (20112)
AC	18248 (19804)	18661 + (17289)	19532 - (19920)	19806 - (20211)	19814 - (20258)
TG	18216 (19468)	19152 + (17857)	19842 (19841)	20108 (20070)	20141 (20126)
TC	11910 (19603)	10663 - (17365)	11580 - (20002)	11739 - (20277)	11782 - (20262)
Trinucleotide motifs					
TAA	60 - (161)	532 - (675)	100 - (270)	101 - (278)	103 - (281)
TAT	64 - (159)	563 - (679)	116 - (270)	116 - (273)	115 - (281)
TAG	541 - (740)	867 - (1225)	561 - (943)	579 - (949)	577 - (932)
TAC	582 - (748)	843 - (1195)	531 - (932)	531 - (952)	547 - (932)
GAA	1599 + (742)	1294 + (1221)	1394 + (931)	1416 + (939)	1417 + (950)
GAT	2452 + (746)	2244 + (1222)	2276 + (929)	2317 + (946)	2324 + (940)
GAG	2113 - (3431)	1083 - (2229)	1449 - (3109)	1472 - (3181)	1471 - (3196)
GAC	9822 + (3458)	3881 + (2188)	7253 + (3137)	7320 + (3194)	7284 + (3208)
GTT	1508 + (750)	2440 + (1239)	2040 + (944)	2067 + (940)	2060 + (952)
GTG	7462 + (3397)	4450 + (2269)	6322 + (3127)	6397 + (3153)	6397 + (3197)
GTC	9701 + (3430)	4079 + (2213)	7406 + (3134)	7504 + (3194)	7484 + (3214)
CAA	1571 + (742)	2325 + (1178)	1863 + (942)	1885 + (938)	1896 + (964)
CAT	2583 + (751)	2108 + (1180)	2263 + (952)	2299 + (947)	2296 + (961)
CAG	7757 + (3420)	4144 + (2198)	5424 + (3146)	5484 + (3165)	5491 + (3221)
CAC	7393 + (3482)	4157 + (2131)	6383 + (3210)	6458 + (3234)	6459 + (3229)
CTT	1615 + (736)	1265 (1232)	1431 + (943)	1443 + (944)	1449 + (947)
CTG	7786 + (3460)	4005 + (2220)	5510 + (3154)	5616 + (3184)	5588 + (3170)
CTC	2079 - (3468)	1030 - (2165)	1624 - (3189)	1648 - (3236)	1647 - (3223)
CGG	23186 + (15900)	5532 + (3989)	16984 + (10607)	17199 + (10710)	17311 + (10778)
CGC	23973 + (15916)	5313 + (3915)	17639 + (10657)	17712 + (10824)	17751 + (10851)

Furthermore, mycobacterial genomes generally show scarcity of tract densities of long microsatellites tracts. In fact the severity of the scarcity of tracts increases proportionately with the increase in the number of repeats. We also calculated the observed/expected (O/E) ratios of microsatellites with different repeat numbers in the genomes of *E. coli* K12 that consists of all the post-replicative DNA repair enzymes and *H. pylori* (both the strains: 26695 and J99) that possesses only a partial repair system marked by the absence of enzymes *mutL* and *mutH* (Figure 3.2)(Tomb et al. 1997).

Comparatively the O/E ratio of microsatellites with higher repeat numbers is lower in *E. coli* as compared to the *H. pylori* and hence this feature correlates with the regulatory role of repair systems to control long repeat tracts. However, it is surprising to note that mycobacterial genomes are represented by the lowest O/E ratios of the long repeats. This observation is quite contrary to what one might believe that the complete absence of the mismatch repair system would enrich a genome with long repeat tracts. Hence, it is certain that selection plays a paramount role in imposing restriction on the expansion of microsatellite tracts in mycobacterial genomes.

3.3.3 Potential polymorphic microsatellites (PPMs)

Mutations in microsatellites are believed to be dependent on their tract lengths; long tracts with more than seven repeats are more prone to slippage than shorter tracts (Brinkmann et al. 1998) and hence any such tract can be called as a potential polymorphic microsatellite (PPM). Although mycobacterial genomes show scarcity of long tracts, a few tracts which could be classified as PPMs are found. It is worthwhile to examine the locations of these tracts as microsatellite polymorphism in the coding regions can bring about in-frame or out-of-frame mutations, while in non-coding regions can affect regulatory signals situated upstream of the coding regions (Gur-Arie et al. 2000).

PPMs in the non-coding regions

Among the mycobacterial genomes, *M. avium* is completely devoid of PPMs in its non-coding regions. *M. bovis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv harbor some PPMs, while *M. leprae* is relatively richer in PPMs with thirty-eight tracts (Table 3.12). Most of the PPMs (38 out of 44) found in these genomes comprise of mononucleotide tracts. Of the 38 PPMs, majority of them are situated farther than 200bp away from their upstream coding regions, thereby hinting no effect of their polymorphism on the regulatory elements of the downstream coding regions. In *M. leprae* the PPM is a dinucleotide tract (AT)8 and is located 79bp away from an ORF annotated as a*tetR* (tetracycline resistance) family transcriptional regulator. This gene encodes for a repressor protein that regulates the expression of the membrane associated protein *tetA* protein involved in the export of tetracycline out of the bacterial cell (Hinrichs et al. 1994; Kisker et al. 1995). Five of these thirty-eight PPMs from *M. leprae* have already been tested and reported to be polymorphic in some clinical isolates (Table 3.11) (Groathouse et al. 2004). All these variable microsatellites are 200bp away from the coding regions. These microsatellites are used as molecular markers for strain typing (Groathouse et al. 2004). In *M. tuberculosis* CDC1551, one of the two PPMs viz., C_8 is 29bp away from the HIT (histidine triad) family protein. The function of this family of proteins is unknown, however, they are conserved in various prokaryotes as well as eukaryotes (Seraphin, 1992). The PPMs in *M. tuberculosis* H37Rv and *M. bovis* seem to be the equivalents of the PPMs in *M. tuberculosis* CDC1551, but the downstream coding regions have been annotated as hypothetical proteins.

PPMs in the coding regions

Among the PPMs found in the coding regions, 13 are present in *M. leprae*, 9 in *M. avium*, 10 in *M. bovis*, and *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv harbor respectively 5 and 6 PPMs (Table 3.12). Interestingly, all the PPMs are

mononucleotide tracts, and therefore insertion or deletion of mono repeat units (unless there is a simultaneous insertion or deletion of three mononucleotides repeats or their integral multiples) leads to shifts in the reading frame causing either premature terminations or new translated sequences. In all the genomes PPMs are distributed in the ORFs encoding non-house keeping genes such as membrane proteins, virulence factors, PPE proteins, as well as hypothetical proteins.

3.4 Discussion

As can be anticipated, the distributions of microsatellites in the closely related *M. bovis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv genomes, are similar to each other. *M. leprae* and *M. avium* show distinct distribution profiles. Although SSRs are distributed throughout the mycobacterial genomes, there are some regions that are markedly either rich or poor in them. *M. bovis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv have more number of such microsatellite rich regions than the poor regions. Many stress response genes, transcription regulators and virulence factors are embedded in the repeat rich regions. Genes that are unique to mycobacteria, such as PE and PPE are present in repeat rich regions. Hence, it appears that the repeat rich regions act as reservoirs for genes, that are capable of bringing about certain variability in virulence, antigenicity and host adaptation. In stressful conditions, increased microsatellite mutations could generate gene variants in different populations, thus conferring a stress response to tolerate and survive in hostile environments. The presence of a large number of repeat enriched regions in *M. bovis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv as compared to *M. avium* and *M. leprae*, is perhaps a reflection of the exposure to hostile environments that these organisms may have to deal with.

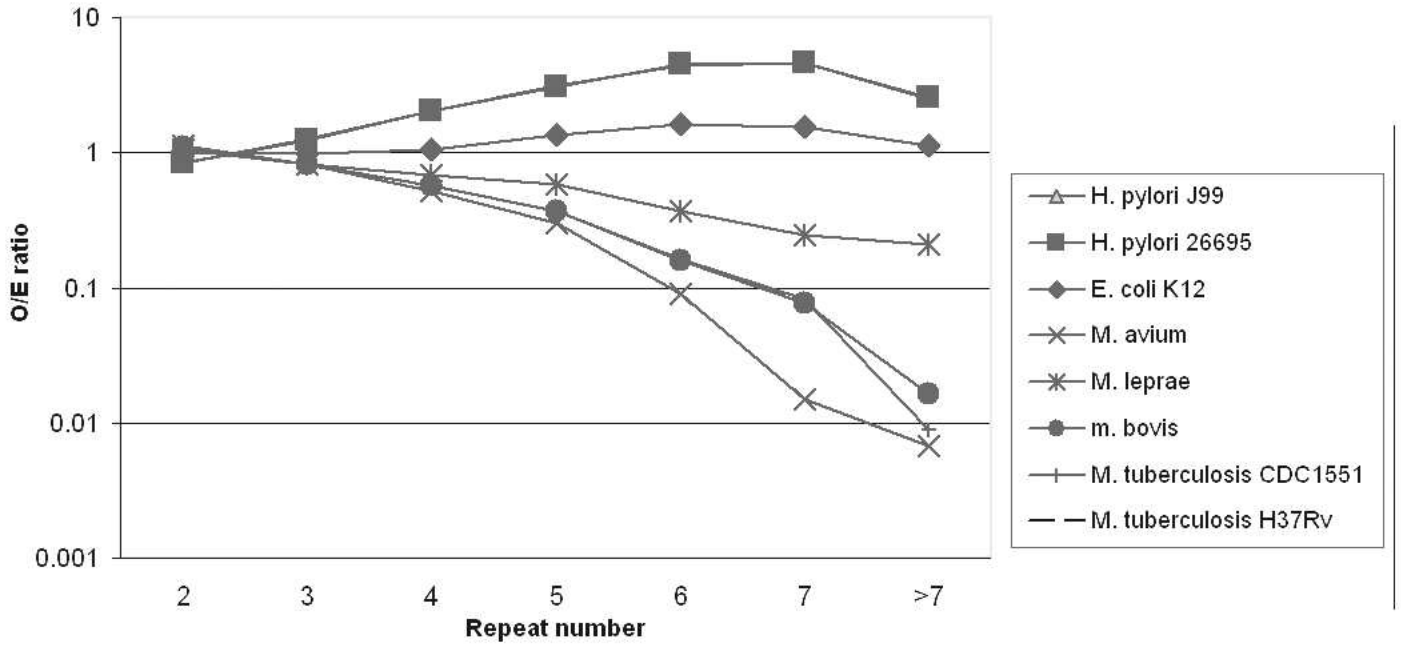

Figure 3.2: Graph showing ratios of observed and expected numbers of microsatellite from *E. coli* K12, *H. pylori* (J99 and 26695), *M. avium*, *M. leprae*, *M. bovis* and *M. tuberculosis* (CDC1551 and H37Rv).

Table 3.12: The list of potential polymorphic microsatellites (PPMs) found in non-coding regions in the five mycobacterial genomes. Repeats which are in **bold** have been tested and reported for their repeat variation (Groathouse et al.(2004)).

Motif	Repeat number	Repeat location in the genome (bp)	Upstream ORF within 200 bp
M. leprae			
T	8	143422	-
C	8	229073	-
G	22	229625	-
C	**20**	**312039**	-
T	8	337466	-
G	10	347280	-
G	10	442993	-
T	8	514193	-
T	8	634418	-
G	8	663135	-
G	8	667968	-
G	8	741133	-
G	8	755942	-
G	9	976857	-
G	8	1197267	-
G	11	1309544	-
A	9	1414666	-
C	8	1778021	-
C	16	1987156	-
G	8	1987172	-
T	8	2486597	-
A	8	2562329	-
C	9	2658192	-
A	8	2946873	-
T	8	3215279	-
AT	14	308814	-
AT	15	948935	-
TA	**18**	**984591**	-
AC	9	1452573	-
CA	8	1531184	-
TA	10	1744091	-
AC	8	2211035	-
AT	**17**	**2597735**	-
AT	10	2844970	-
TA	11	2951820	-

(continued)

Table 3.12: (continued...)

Motif	Repeat number	Repeat location in the genome (bp)	Upstream ORF within 200 bp
AT	8	3221616	Putative TetR-family transcriptional regulator (GI:15828443)
GTA	**9**	**2583814**	-
GAA	**21**	**2785433**	-
M. bovis			
C	8	856443	Hyothetical protein (GI:31791947)
G	27	1619414	-
G	8	4036749	-
M. tuberculosis CDC1551			
C	8	191560	-
C	8	856394	HIT family protein (GI:15840174)
M. tuberculosis H37Rv			
C	9	854251	Hypothetical protein Rv0759c (GI:15607899)

Table 3.13: The list of potential polymorphic microsatellites (PPMs) found in coding regions in the five mycobacterial genomes.

Motif	Repeat number	Repeat location in the genome (bp)	Gene name
M. avium			
C	8	169357	Hypothetical protein (41406264)
C	19	1793090	Hypothetical protein (41407736)
C	8	2119844	Hypothetical protein (41408016)
G	8	2467982	Hypothetical protein (41408318)
C	10	2719084	Hypothetical protein (41408519)
C	8	2820932	Hypothetical protein (41408610)
C	8	3300015	Hypothetical protein (41409061)
G	8	3880098	GuaA (41409587)
G	8	4209441	Hypothetical protein (41409863)
M. leprae			
C	8	22194	Putative penicillin-binding protein (15826881)
C	8	67158	Putative membrane protein (15826903)
C	8	151870	Possible membrane protein (15826944)
T	8	170900	Hypothetical protein (15826958)
A	8	225690	Putative phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase (15826983)
G	8	299803	Putative membrane protein (15827025)
A	8	593037	Putative protein-export membrane protein (15827165)
T	8	795734	Hypothetical protein (15827271)
A	8	893789	Putative dTDP-rhamnose modification protein (15827318)
G	12	1116443	Conserved hypothetical protein (15827450)
G	8	1145474	Cell division protein (15827463)
T	8	1511004	Putative antiporter (15827650)
G	8	3093597	Putative cell invasion protein (15828393)
M. bovis			
G	8	17896	Serine/threonine protein kinase (31791192)
C	8	693132	MCE-family protein MCE2DA [FIRST PART] (31791774)
G	9	977363	PPE family protein (31792066)

(continued)

Table 3.13: (continued...)

Motif	Repeat number	Repeat location in the genome (bp)	Gene name
G	8	1168426	Hypothetical protein (31792236)
C	8	1543144	Glucolipid sulfotransferase [first part] (31792567)
G	11	1744180	frdB and frdC (31792738)
C	11	2320081	Transmembrane protein (31793264)
C	15	2771732	PE-PGRS family protein [first part] (31793670)
C	8	3321471	Lipoprotein LPPZ (31794183)
C	8	4076531	GLPKA (31794867)
M. tuberculosis CDC1551			
G	8	17897	Serine/threonine protein kinase (15839389)
T	8	976902	PPE family protein (15840292)
G	9	976910	PPE family protein (15840292)
C	8	2340527	Hypothetical protein (15841574)
C	8	3359231	Lipoprotein, putative (15842564)
M. tuberculosis H37Rv			
G	8	17897	PknA (15607157)
T	8	976888	PPE (15608018)
G	9	976896	PPE (15608018)
C	8	1992322	PE_PGRS(wag22) (15608897)
C	8	2338193	Hypothetical protein Rv2081c (15609218)
C	8	3364853	LppZ (15610143)

By and large, microsatellite motif distributions are similar to those found in other prokaryotes. Under-representation of mononucleotide, di, tetra and penta repeats is commonly observed in many prokaryotic genomes (Field and Wills, 1998) along with the over-representation of the trinucleotide and hexanucleotide repeats. Under-representation of di, tetra and penta motifs in the genomes where most parts are coding, can be attributed to the existence of selection pressures to avoid chances of frameshift mutations brought out by these microsatellites in the coding regions. Therefore, in *M. leprae* where nearly half the genome is non-coding, a less selection pressure against frameshift mutations is expected. Indeed our analysis indicates that the tetra and penta repeats are excessively represented in this genome.

Among the microsatellites, the mono, di and tri with iterations of two are in excess and such abundance has also been reported in some of the prokaryote genomes (Field and Wills, 1998). The codon usage has been attributed to such excess of short repeat sequences (Field and Wills, 1998).

In all the mycobacterial genomes the base composition of the genomes appears to be influencing the abundance as well as the enrichment of microsatellites. Most of the di to hexa tracts are G+C rich and their numbers are in excess. The mono tracts show different characteristics wherein the most frequent G/C tracts are under-represented, while the less abundant A/T tracts are over-represented. This indicates a trend in the evolution of mononucleotide tracts, wherein A/T tracts seem to get accumulated. Among the dinucleotide tracts under-representation of the TA repeats is observed in many prokaryotic genomes, and this repeat has been considered as universally under-represented (Burge et al. 1992; Karlin et al. 1997). TA forms a part of regulatory sequences, and therefore its depletion in genomes has been attributed to the avoidance of inappropriate binding of the regulatory elements (Burge et al. 1992).

Mutations in SSRs especially, small motifs (mono and dinucleotide) with high repeat number are more prone to mutations than long motifs with low re-

peat number (Shinde et al. 2003). In the studied genomes, all the long repeats (PPMs) located in the coding regions are mononucleotide repeats only, hinting that these coding regions may act as contingency loci. Most of the contingency genes in pathogenic bacteria code for membrane proteins and membrane associated proteins (Moxon et al. 1994) favoring antigenic variation, thus conferring a particular selective advantage to escape the host immune system.

The distribution of microsatellites in a genome is considered as an equilibrium between the expansion due to addition of repeat units and the point mutations that break long microsatellites into smaller tracts (Kruglyak et al. 1998). The length polymorphism of a repeat tract is primarily controlled by the selection forces that act on it (Nauta and Weissing, 1996). Hence, microsatellite distribution and frequency in a genome reflects the underlying mutational processes, selection constrains as well as DNA repair mechanisms. In all the mycobacterial genomes, there seems to be a strong control on the tract lengths as compared to other bacterial genomes such as *E. coli*. In fact the per Kbp distribution of repeats in mycobacterial genomes is in the range of 220-230 which is equivalent to that observed in *E. coli* and *H. pylori* (data taken from MICdb, (Sreenu et al. 2003)), thus indicating a tight regulation of microsatellite evolution (birth, mutation and death) in spite of the deficiency of a post replicative repair mechanism. These observations strongly suggest the control of microsatellite evolution in mycobacterial genomes by various selection mechanisms.

3.5 Summary

In this chapter we presented an analysis of perfect microsatellites with reference to their distribution and abundance in the five, fully sequenced, mycobacterial genomes viz., *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv, *M. bovis*, *M. leprae* and *M. avium*. It was found that the distribution profiles of microsatellites in the closely related genomes *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv

and *M. bovis*, are very similar. All the genomes, except *M. leprae*, possess some regions which are characteristically enriched and deficient in microsatellites. The most interesting discovery of our studies is that these genomes have a general scarcity of long microsatellite tracts despite the absence of post-replicative repair system, indicating strong role of selection against expansions of microsatellites in these genomes. However, sparingly found long microsatellites can be considered as potential polymorphic sites which can impart some plasticity to these genomes.

Chapter 4

Microsatellite polymorphism across *M. tuberculosis* and *M. bovis* genomes: Implications on genome evolution and plasticity

---

Publications from this chapter

- **Sreenu, V.B.**, Kumar, P., Nagaraju, J. and Nagarajaram, H. A. (2005) Microsatellite polymorphism across mycobacterial genomes: Implications on genome plasticity and adaptability (communicated).

4.1 Introduction

Pathogenic bacteria encounter hostile environments while gaining access to host, during colonization and proliferation. In order to survive and confront selections pathogens maintain sufficient degree of genetic plasticity. Mutations in the bacteria can create rare variants which can survive hostile conditions. Polymorphic microsatellites too are known to provide crucial genetic plasticity with their reversible frameshift mutations which can affect transcription as well as translation of genes in the form of regulation of expression levels (Moxon et al. 1994), switching on/off genes (Moxon et al. 1994) and altering the gene functions (Ritz et al. 2001).

In the previous chapter it was shown that the mycobacterial genomes harbor a number of microsatellites of which some can be considered as potential polymorphic sites as their tract lengths are above the threshold which is generally believed to be the minimum length for a tract to undergo slippage. It is interesting to analyze these genomes for polymorphic microsatellites and study their role in providing genetic variability to the pathogens.

Earlier reports on genomic changes in *M. tuberculosis* were mainly concerned with single nucleotide polymorphisms (SNPs) and large-sequence polymorphisms (LSPs) (>10 bp) (Fleischmann et al. 2002). Some of these varying sequences are located in the genes involved in host pathogen interactions where the changes allow the bacteria to better adaptations. Involvement of SNPs in the drug resistance is also shown previously in this genome (Blanchard, 1996). However, to date there is no report available on microsatellite mutations. In the present study, we show for the first time that the coding regions of the three genomes of mycobacteria (the two strains of *M. tuberculosis*, [H37Rv(Cole et al. 1998) and CDC1551(Fleischmann et al. 2002)] as well as *M. bovis* (Garnier et al. 2003)) harbor a number of polymorphic microsatellite loci which show polymorphism and such polymorphic microsatellites are associated with remarkable changes to

the corresponding coding regions.

4.2 Methods

Complete genome sequences of *M. tuberculosis* (H37Rv and CDC1551) and *M. bovis* were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The microsatellites in the genomes were identified using SSRF (Sreenu et al. 2003). ORFs harboring microsatellites from one genome were used as queries and searched against the entire genome sequence of another organism using locally installed BLASTN program (version 2.2.6) (Altschul et al. 1997). Indexing of the database (complete genome sequence) was done with the default parameters. Repeat masking filter (-F F) and a 100bp dropoff value for final gapped alignment (-Z 100) were used along with the default BLASTN search parameters. The hits that gave rise to alignments with queried sequences with indels only in the microsatellites, were selected for further analysis. These regions (can be coding or non-coding in the second genome) were realigned using CLUSTALW (Thompson et al. 1994) to reconfirm the alignment as well as the indels in the microsatellites. For functional annotations of the coding regions references were made to tuberculist website (http://genolist.pasteur.fr/TubercuList/) and the TB structural genomics consortium site (http://www.tbgenomics.org).

4.3 Results

4.3.1 Polymorphic microsatellites

Table 4.1 gives the number of polymorphic microsatellites found from the pairwise comparison of the three genomes. The complete list of the polymorphic microsatellites and the coding regions harboring them are given in Tables 4.2. A quick glance at this table yields the following points:

- Most of the microsatellites which have undergone polymorphism are made up of mononucleotide repeats
- Majority of the polymorphisms have caused frameshift mutations in the coding regions resulting in a variety of changes to the coding regons
- Even the short microsatellites of repeat number 2 have undergone length variation due to indels of repeat units
- Of the genome pairs, *M. tuberculosis* CDC1551 - *M. bovis* and *M. tuberculosis* CDC1551 - *M. tuberculosis* H37Rv have recorded maximum and minimum number of polymorphic microsatellites respectively
- A large majority of the tracts are G/C tracts

An indepth analysis of the polymorphic microsatellites yielded the following information. Indels of repeat units have caused frameshifts in many ORFs harboring them, which in turn have brought out interesting changes to those ORFs as illustrated schematically in Fig.4.1. In some cases the ORFs have been split (fission),[1] in some others two adjacent ORFs have been fused with or without overlap to form one single ORF and there are also cases where the ORFs have either been eliminated due to premature termination by stop codons, or undergone changes in their lengths.

ORF Fusion/fission

We have carried out a detailed analysis of the genes that have undergone fusion or split due to mycrosatellite polymorphism and the results are discussed separately in Chapter 5. However, for the sake of continuity we summarize some of the highlights in the following paragraph.

[1]We define split event as the one where one ORF (annotated as one functional protein) has split into two ORFs (annotated as two parts: protein a and protein b) in the reference genome and fusion as the one where two ORFs (annotated as two functional proteins) have fused into one ORF in the genome of reference.

Table 4.1: Number of polymorphic microsatellites found from the pair-wise comparison

M. tuberculosis H37Rv - *M. bovis*	58
M. tuberculosis H37Rv - *M. tuberculosis* CDC1551	32
M. tuberculosis CDC1551 - *M. bovis*	74

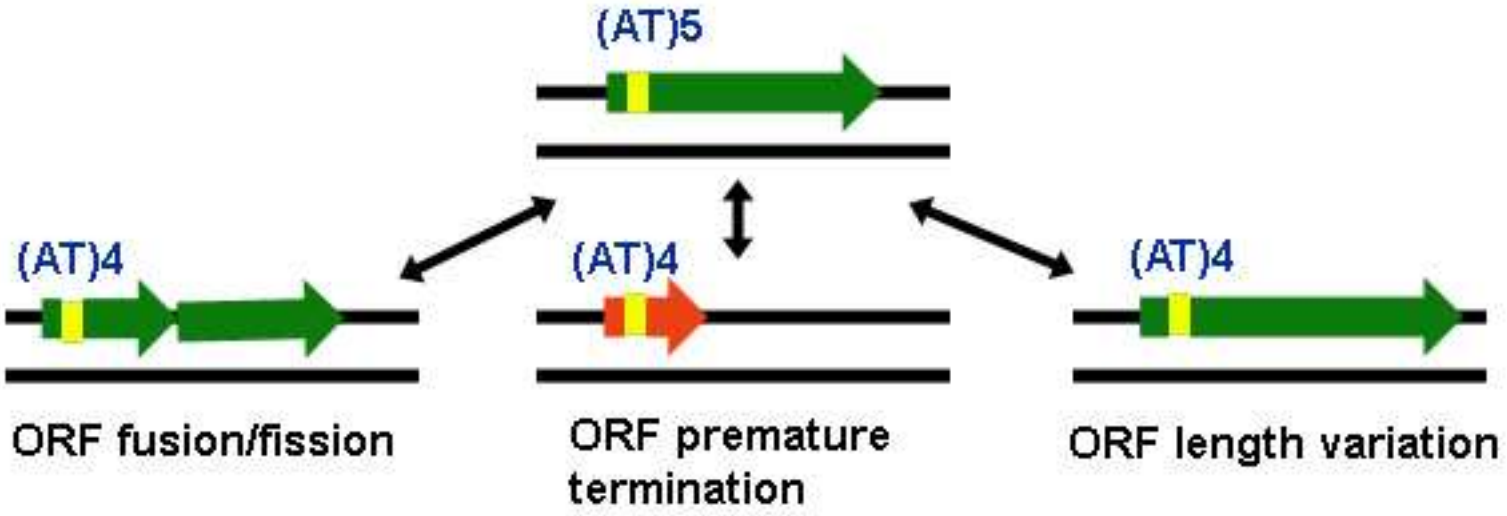

Figure 4.1: A schematic representation of the various effects caused by indels of repeat units in microsatellites in the coding regions (shown as arrows) of the three mycobacterial genomes, *M. bovis*, *M. tuberculosis* (CDC1551) and *M. tuberculosis* (H37Rv).

Of the three genomes *M. bovis* has accommodated a large number of split events. All the ORFs that have split in *M. bovis* are those that have been annotated as membrane proteins, transporters, PE_PGRS, cell-wall synthesis proteins and hypothetical proteins. Membrane proteins are known to play an important role in host-pathogen interactions (Stern and Meyer; 1987) and many bacteria are known to change their membrane protein structures to escape host immune defense system and to colonize in other places in the host (Moxon et al. 1994; Stern and Meyer, 1987). The PE_PGRS proteins are specific to mycobacteria and have been speculated to function as surface antigens (Banu et al. 2002; Brennan et al. 2001). It has to be noted that in the case of split ORF, expression of the second part of the ORF is entirely dependent on the regulatory mechanism of that ORF and the second part may not even get expressed. If the ORF is coding for a surface protein then the truncation would lead to antigenic variation. One could surmise that, split ORFs in *M. bovis* genome implicate a scope for versatile protein-protein interactions provided both the parts get expressed. If they are expressing the two subunits can act together (Enright et al. 1999; Marcotte et al. 1999) as well as get involved in other protein-protein interactions and this might create a favourable situation for operating alternate/new pathways which in turn may eventually render better adaptability to the bacteria. This perhaps could be the probable reason for *M. bovis* to have a wider host range as compared to *M. tuberculosis*.

Table 4.2: List of ORFs (given by their gene id along with number of amino acids in parentheses) from mycobacterial genomes harboring polymorphic microsatellites. The first column depicts microsatellite tract and its observed mutation in the form of insertion/deletion of repeat units leading to expansion or contraction of the microsatellite. The observed event being a case of insertion or deletion of repeat is decided by the number of genomes in which the repeat number is conserved (given as bold text). For example, G**4**⇒5 means in two of the genomes the tract is G4 and in the third genome it is G5 and therefore it is an event of insertion leading to microsatellite expansion.

Mutation	ORF's function	*M. tb* H37Rv	*M. tb* CDC1551	*M. bovis*
Muation leading to premature termination				
C**7**⇒8	Oxido-reductase	**Rv0161 (449)**	MT0170 (pseudo)	**Mb0166 (449)**
T**5**⇒4	*uma*A1	**Rv0469 (286)**	MT0485 (pseudo)	**Mb0478 (286)**
G**4**⇒3	Cysteine synthase	**Rv0848 (372)**	MT0871 (pseudo)	**Mb0871 (372)**
G**3**⇒2	Membrane transport	**Rv0849 (419)**	MT0872 (pseudo)	**Mb0872 (419)**
A**2**⇒3	Hypothetical protein	-	**MT1025.1 (46)**	-
G**4**⇒3	Polyketide synthase *pks*5	**Rv1527c (2108)**	Pseudo gene	**Mb1554c (2108)**
G**3**⇒2	Conserved hypothetical	**Rv1533 (375)**	Pseudo gene	**31792719 (375)**
G**7**⇒8	PE_PGRS(wag22) Antigen	Rv1759c (914)	**MT1807 (pseudo)**	**Mb1789c (820)**
G**3**⇒2	PE_PGRS	**Rv2126c (256)**	MT2185 (pseudo)	**Mb2150c (256)**
G**2**⇒3	Hypothetical protein	Not annotated as orf	**MT2401.2 (69)**	pseudo
CGCGC**2**⇒3	Oxidoreductase	**Rv3093c (334)**	MT3177 (pseudo)	**Mb3120c (334)**
A**3**⇒2	Conserved hypothetical protein	Pseudo	**MT3855 (314)**	**Not annotated as orf**
G**3**⇒2	MycP2, membrane-anchored serine protease	**Rv3886c (550)**	MT4001 (pseudo)	**Mb3916c (550)**
C**7**⇒8	Glycolipid sulfotransferase	**Rv1373 (326)**	**MT1418 (320)**	Mb1407 (265)
C**6**⇒5	Hypothetical protein	**Rv1718 (272)**	**MT1757 (386)**	Mb1746 (207)
G**7**⇒8	*glp*K, glycerol kinase	**Rv3696c (517)**	**MT3798 (517)**	Mb3721c
C**2**⇒3	*sigM*	Rv3911 (222)	**MT4030 (196)**	**Mb3941 (196)**
Mutation leading to length variation				
T**5**⇒4	*ctpI*	Rv0107c(1632)	**MT0116 (1625)**	**Mb0111c (1625)**

(continued)

Table 4.2: (continued...)

G**4**⇒3	Hypothetical protein	**Rv0607 (128)**	MT0636 (147)	**Mb0623 (128)**
G**3**⇒2	*lldD1*	**Rv0694 (396)**	MT0721 (419)	**Mb0713 (396)**
C**4**⇒5	*nusB*	**Rv2533c(156)**	MT2608 (290)	**Mb2562c (156)**
G**3**⇒2	Transport proteins	**Rv3239c (1048)**	MT3337(1065)	**Mb3267c (1048)**
GC**4**⇒5	Hypothetical protein	**Rv0739 (268)**	**MT0764 (268)**	Mb0760 (282)
A**3**⇒2	Hypothetical protein	Rv1046c (174)	**MT1075.1 (262)**	**Mb1075c (197)**
C**5**⇒4	Conserved hypothetical protein	**Rv1760 (502)**	**MT1809 (531)**	Mb1791 (509)
GC**2**⇒1	*hflX*	**Rv2725c (495)**	MT2797 (556)	**Mb2744c (495)**
T**4**⇒3	Integral membrane protein	**Rv3162c (145)**	**MT3251 (145)**	Mb3187c (196)
C**3**⇒2	ESAT-6 like protein	**Rv3890c (95)**	**MT4005 (95)**	Mb3919c (124)
G**3**⇒2	Conserved hypothetical protein	**Rv1246c (97)**	MT1284 (143)	**Mb1278c (97)**
AC**5**⇒6	*lprJ*	**Rv1690 (127)**	**MT1729 (127)**	Mb1716 (139)
G**5**⇒4	Conserve membrane protein	**Rv3693 (440)**	MT3795 (475)	**Mb3718 (440)**
G**2**⇒3	PBP-4 (penicilline binding)	**Rv0907(532)**	MT0930 (562)	**Mb0931 (516)**
G**6**⇒5	*moac2*	**Rv0864 (167)**	**MT0887 (167)**	Mb0888 (142)
T**3**⇒2	Membrane protein	**Rv1101c (385)**	**MT1133 (385)**	Mb1131c (342)
A**6**⇒5	*aroE*	**Rv2552c (269)**	**MT2629 (269)**	Mb2582c (260)
G**2**⇒3	Membrane protein	Rv2732c (204)	MT2802.1(180)	Mb2791c (204)
G**3**⇒2	Conserve membrane protein	**Rv3885c(537)**	MT4000 (422)	**Mb3915c (537)**
G**6**⇒5	Membrane protein	**Rv0010c (141)**	**MT0013 (141)**	Mb0010c (111)
A**3**⇒2	Conserved hypothetical protein	**Rv0025 (120)**	MT0028 (90)	**Mb0026 (120)**
C**3**⇒2	NLP/P60 Antigen	**Rv0024 (281)**	**MT0027 (281)**	Mb0024 (277)
C**7**⇒8	*mce2D*	**Rv0592 (508)**	**MT0622 (508)**	Mb0607 (478)
A**8**⇒7	PPE	**Rv0878c (443)**	**MT0901 (444)**	Mb0902 (438)
C**5**⇒6	PPE	**Rv1168c (346)**	**MT1205 (346)**	Mb1201c (180)
CG**5**⇒4	Secretory protein	**Rv1312 (147)**	**MT1352 (147)**	Mb1344 (144)
G**4**⇒3	Hypothetical protein	**Rv1725c (236)**	MT1766 (187)	**Mb1754c (236)**
TG**2**⇒1	*sseB*	**Rv2291 (284)**	**MT2348 (268)**	Mb2314 (256)

(continued)

Table 4.2: (continued...)

G**2**⇒3	UDP-glucosyltransferases	**Rv2958c (428)**	**MT3034 (428)**	Mb2982c (366)
G**3**⇒2	Cyclase	**Rv3377c (501)**	**MT3487 (501)**	Mb3411c (483)
G**2**⇒3	Conserve hypothetical	**Rv3836 (137)**	**MT3944 (133)**	Mb3886 (116)
CGGCCC**1**⇒2	Lipoprotein	**Rv0838 (256)**	**MT0860 (231)**	Mb0861 (258)
GGC**5**⇒4	PE-PGRS	Rv0872c (606)	**MT0894 (609)**	**Mb0896c (608)**
CGG**5**⇒4	PPE	**Rv2356c (615)**	**MT2425 (615)**	Mb2377c (614)
GCC**4**⇒3	PE_PGRS	**Rv2396 (361)**	**MT2467.1 (382)**	Mb2418 (360)
TCGACG**1**⇒2	Hypothetical protein	**Rv1434 (45)**	MT1478 (47)	**Mb1469 (45)**
G**8**⇒11	Membrane protein	**Rv2081c (146)**	**MT2143 (150)**	Mb2107c (147)
GGC**4**⇒3	*gdh*	**Rv2476c (1624)**	**MT2551 (1624)**	Mb2503c (1623)
G**6**⇒3	Transcription regulatory	**Rv2621c (224)**	**MT2696 (224)**	Mb2654c (223)
GCG**5**⇒4	PPE	**Rv3159c (590)**	**MT3247 (603)**	Mb3183c (589)
TGG**4**⇒5	Membrane protein	**Rv2799 (209)**	**MT2867.1 (209)**	Mb2822 (210)
CCG**4**⇒3	*moeZ*	**Rv3206c (392)**	**MT3301 (392)**	Mb3231c (391)

Premature termination

M. tuberculosis CDC1551 strain, as compared to the other genomes, exhibits most of the cases of premature terminations (10 ORFs) (see Table 4.2) and are confined to the PE_PGRS, *umaA1, pks5* and six hypothetical proteins. It is interesting to note that *umaA1* codes for a mycolic acid methyl transferase that modifys lipids of the mycobacterial cell wall (Glickman et al. 2001) and plays an important role in *M. tuberculosis* virulence. It has been shown in mouse model that *umaA1* deletion mutant is more virulent than the wildtype (McAdam et al. 2002). *pks5* codes for a polyketide synthase which is used in the synthesis of secondary metabolites called polyketides. It has been shown in *M. tuberculosis* that mutations in *pks5* gene render severe growth defects to the pathogen (Rousseau et al. 2003).

Length variation

From the three genomes 43 ORFs have undergone length variation due to microsatellite mutations (see Table 4.2). Many proteins in this category have been annotated as hypothetical proteins, PPE and mammalian cell entry (mce) family virulence proteins. In some cases there seems to be no effect on the function of the translated product is seen as the functional domains are conserved whereas in other cases drastic changes are observed. For example, Rv2732c in *M. tuberculosis* H37Rv as well as Mb2791c in *M. bovis* code for a membrane anchoring protein of length 204 aa. The equivalent ORF MT2802.1 in *M. tuberculosis* CDC1551 codes for a shorter protein of 180 aa due to single G insertion in the microsatellite tract GG. In silico analysis of both the proteins revealed that only the shorter protein with N-terminal deletion of 24 aa had high probability (0.959, from signalP (http://www.cbs.dtu.dk/services/SignalP/)) for signal peptide at its N-terminal end and therefore has a potential to secrete whereas the longer protein showed very negligible propensity for signal peptide and probably does not have a potential to secreted out.

4.4 Discussion

Although the three genomes show shortage of long microsatellites (this aspect has already been discussed in the previous chapter), there is no dearth for polymorphic events as, unexpectedly, even the short microsatellites show polymorphism. Microsatellites with as few as two repeats have undergone indels of repeat units (Please see table 4.2). This is in contrast to the earlier observations that invoke requirement of minimum microsatellite length for repeat expansion or contraction due to strand slippage.(Dechering et al. 1998; Rose and Falush, 1998). From the observations made in this study, it can be suggested that indels of repeats in microsatellites may not entirely be tract length dependent. Whether absence of post replicative DNA mismatch repair system mediated by mutS, mutL and mutH genes (Springer et al. 2004) has any direct or indirect role in such unexpected expansions/contractions of short repeats, is a topic for further investigations. Whatever may the underlying molecular mechanism, polymorphic microsatellites are associated with changes to the coding regions which in turn account for a certain amount of genomic plasticity. However, how such changes are beneficial to the adaptability, virulence and survival of these pathogens, is worth an investigation. For this, the polymorphic microsatellites discovered in this study can form a good starting set to be screened for mutations in large number of clinical isolates. It is also quite probable that some of the polymorphic microsatellites discovered in this study are simply the molecular fossils, which might have helped to nucleate necessary diversity to give rise to speciation.

4.5 Summary

We show, for the first time, that the mycobacterial genomes harbour number of polymorphic microsatellites in the coding regions despite the shortage of long microsatellites. Between any two mycobacterial genomes, repeat number variations is accompanied by striking changes in the corresponding open reading frames

(ORFs). The changes in ORFs observed in between two genomes are: splitting of ORFs, premature terminations, and change of protein function. The observations made in this study throw light on possible roles of microsatellite length polymorphisms in imparting novel functions, thus bringing out plasticity in mycobacterial genomes perhaps associated with molecular mechanisms involved in their adaptability and evolution.

Chapter 5

Microsatellite length variation causes genes to fuse/split in mycobacterial genomes

---

Publications from this chapter

- **Sreenu, V.B.**, Kumar, P. and Nagarajaram, H. A. (2005) Microsatellite length variation causes genes to fuse/split in mycobacterial genomes (communicated).

5.1 Introduction

For any organism, survival is a continuous process often involving reshaping of its genome according to the ever changing environment. In order to survive, new functions are gained and adapted. Altered environments induce an organism to alter its biological activities in order to emerge out as the successful one to thrive in the population (Moxon et al. 1994). It has always been a riddle to scientists as to how organisms acquire new genes to perform completely new functions. With the advent of modern biological research, development of new technologies and availability of complete sequences of whole genomes of different organisms, a platform has now been created to conduct research in order to shed more light on gene acquisition for organism survival.

In many organisms, one of the known ways of procuring completely new genes is through horizontal gene transfer (HGT) (de la Cruz and Davies, 2000; Ochman et al. 2000). Organisms can obtain foreign genetic material readily by transformation, or with the help of viruses and other mobile elements like plasmids and transposons (Garcia-Vallve et al. 2000). Availability of complete genome sequences of several organisms has permitted systematic genome analysis for the presence of HGT and the data obtained have revealed a high rate of HGT in prokaryotes (Garcia-Vallve et al. 2003). Nearly 25% of bacterial genomes account for HGT (Garcia-Vallve et al. 2003). HGT produces extremely dynamic genomes in which a functional gene would be readily gained by bacteria thus acquiring a completely new pathway. However, the occurrence of these transfers are more prevalent in non-pathogenic bacteria than in pathogenic bacteria (Garcia-Vallve et al. 2003; Lawrence and Ochman, 1997).

Apart from acquiring foreign genetic elements, organisms would also tend to use their available genetic material to accomplish the intended task of making changes in the genome. These internal changes could be shifts in the DNA, like

strand inversion, duplications (Gordon and Halliday, 1995; Kant et al. 1985); or translocation (Chambers et al. 2003) or aberrations in the DNA, like point mutations (Ma and Redfield, 2000) or mutations due to small repetitive elements like microsatellites (Moxon et al. 1994). In eukaryotes, exon shuffling also accounts for the gain of new functions without any involvement of foreign genetic material (Patthy, 1996). At the protein level, all DNA shifts bring a new function due to the gain or rearrangement of domains, which otherwise might lead to gene fissions or fusions. These gene fissions and fusions are the major source for the evolution of multi-domain proteins (Ponting and Russell, 2002) and the main cause for the functional evolution of proteins in many organisms. A number of orthologous enzymes of small molecular metabolism in yeast and *E. coli* are shown to have evolved from gene fission and fusion events (Jardine et al. 2002).

Gene fusion/fission could happen due to point mutations in the stop codon or due to frameshift mutations. In this study, we define gene fission as an event when an ORF which is conserved in two of the three genomes is split into two ORFs in the third genome due to a frameshift mutation and the split ORFs are also annotated with the same function as that of the whole ORF in the reference genomes. For example., ORF *aceA* in *M. bovis* has been split into two as *aceA* first part and *aceA* second part in *M. tuberculosis* H37Rv. The ORF fusion is defined as an event where two conserved ORFs adjacent to each other in two of the genomes due to a microsatellite mutation in the first ORF in the third genome makes that ORF to override the stop codon and get fused into the adjacent ORF to give rise to a single ORF (Fig. 5.1).

Fused genes may not always account for fused proteins. As stated by Long (Long, 2000), *An authentic gene fusion should possess a particular mechanism to override the nonsense codon used to stop translation of the N-terminal protein.* Thus, a mutation can make transcription of the first gene to continue without encountering a stop codon and leading to an in-frame fusion with the next gene thus leading to one single mRNA which eventually translates into a fused protein.

Such frameshift mutations can result from spontaneous mutations, or due to microsatellites as shown in this chapter.

In the present study, we reveal the role of polymorphic microsatellites in gene fusion/fission as found from comparative genome sequence analysis of three closely related mycobacterial genomes (*M. tuberculosis* (CDC1551 and H37Rv) *M. bovis*). To the best of our knowledge this is the first ever report of whole genome-wide survey for polymorphic microsatellites involved in gene fusion and fission events in any pathogenic bacteria.

5.2 Methods

The method followed to identify ORFs which have undergone fission/fusion is same as that described in the previous chapter. However, for the sake of continuity we provide a summary of the method below.

Nucleotide sequence of every one of the ORFs harbouring microsatellites from one mycobacterium was compared with the nucleotide sequences of the other two mycobacteria using BLASTN (Altschul et al. 1997). Repeat masking filter (-F F) and a 100bp dropoff value for final gapped alignment (-Z 100) were used along with the default BLASTN search parameters. Of the BLAST hits those showing 100% match within the regions flanking the microsatellite tract which have undergone indels of repeat units with the queried ORFs were selected as the gene loci with polymorphic microsatellites. Of the gene loci only those showing frameshifts were selected for further analysis. Again from the frameshifts those resulting with ORF fusion/fission were recorded.

Paralogous sequences of a translated ORF were identified by querying against all the translated ORFs in the same genome. Hits which were showing more than 30% sequnece identity to complete translated ORF with a similar function were considered as paralogs.

5.3 Results

Comparisons of the three genomes for polymorphic microsatellites affecting the ORFs to undergo fusion/fission, yielded 30 examples. Among them, 27 examples were found affected by indels of mononucleotide repeat tracts and the remaining 3 by indels of dinucleotide repeat tracts. Overall, there were a total of 11 insertions and 19 deletions observed. Indels of repeat units in a microsatellite at the 5' end or in the middle of an ORF was found leading to ORF fission, whereas at the 3' end was found leading to ORF fusion. Expansions/contractions of microsatellites were observed in the short tracts of length as small as 3bp.

5.3.1 Gene fusion

Overall, 9 gene fusion events were observed and all these were found in *M. tuberculosis* CDC1551 (4 events) and *M. bovis* (5 events) brought out by insertion/deletion of repeat units in microsatellites (Table 5.1). The fused ORFs include those encoding hypothetical proteins as well as those assigned known functions. In the following sections we describe in detail the ORFs which have undergone fusion and their possible implications on the ORF function.

In *M. tuberculosis* (both H37Rv and CDC1551), frdB (iron-sulfur protein) and frdC (membrane-spanning domain) known to be involved in anaerobic respiration, are present as two separate proteins and are part of the fumarate operon, with frdB forming a catalytic dimer with frdA, and frdC acting as membrane anchor (Ge et al. 1997; Lauterbach et al. 1990). Among these, frdB harbors the mononucleotide tract G_{11}. In *M. bovis* the G repeat tract has lost four nucleotides due to which the ORF *frdB* has fused into the ORF *frdC*. As suggested by Garnier et al (2003), fusion of these proteins might bring about structural changes by the way of affecting their position in the membrane.

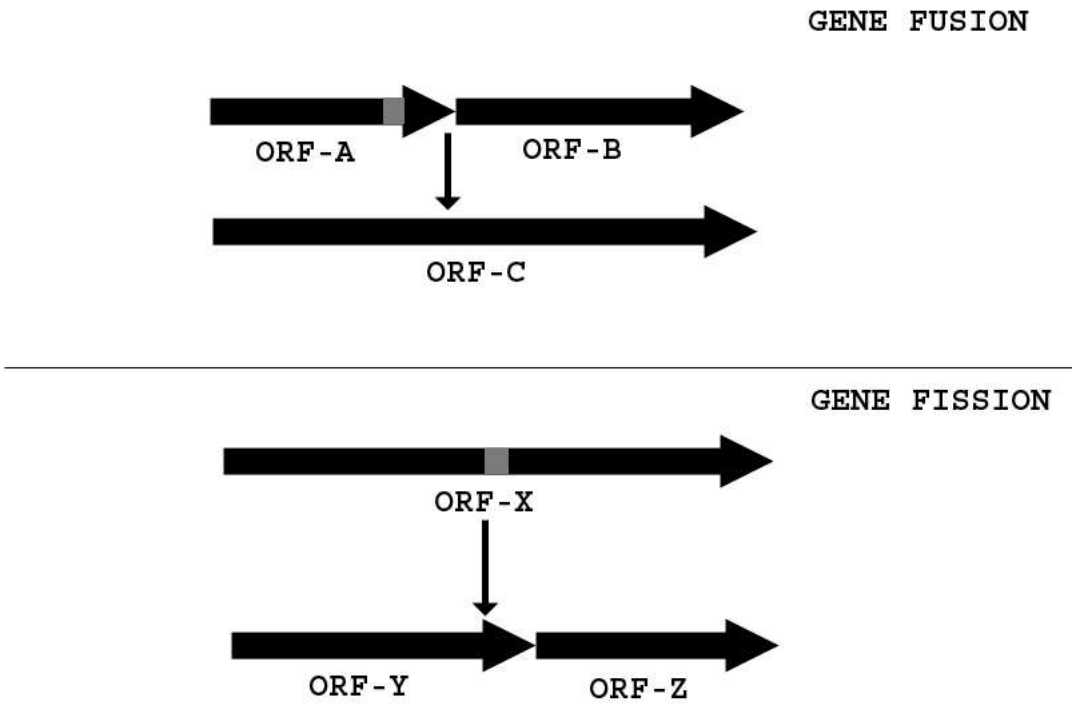

Figure 5.1: Schematic representation of gene fusion and fission. In gene fusion event, mutation in the microsatellite at the end of the ORF-A (shown as gray a band), makes this ORF to fuse to the next ORF (ORF-B) to give rise to a single fused ORF (ORF-C). This fusion does not change the reading frame of the ORF-B (in-frame fusion). In gene fission event, mutation in the microsatellite (shown as a gray band) in the ORF-X results in the genesis of small ORFs (ORF-Y and ORF-Z).

In *M. tuberculosis* CDC1551, *gmhA* that codes for phosphoheptose isomerase is fused to a hypothetical protein due to a deletion of one T in the repeat tract of T_5 (see Table 5.1). Although the functional role of gmhA in *M. tuberculosis* has not been determined, its homolog is known to be involved in lipooligosaccharide biosynthesis and has a key role in virulence in *H. influenzae* (Bauer et al. 1998). We also analyzed the hypothetical protein for its probable function using DOMAIN FISHIN (Contreras-Moreira and Bates, 2002) (http://www.bmm.icnet.uk/servers/3djigsaw/dom_fish) and NCBI conserved domain database search (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cdd) (Marchler-Bauer et al. 2005) and found that it contains histidinol phosphatase domain (HisB). Even though hisB and gmhA are functionally do not seemed to be linked, addition of HisB domain to gmhA may bring some changes in the virulence of the pathogen. However, another copy of the *hisB* gene, that codes for imidazole glycerol-phosphate dehydratase is also present in *M. tuberculosis* CDC1551.

rfbB is a gene involved in the conversion of dTDP-d-glucose into dTDP-l-rhamnose, which acts as a donor for L-rhamnose, one of the molecules found in the cell wall of *M. tuberculosis* and other pathogens (Dong et al. 2003). In *M. tuberculosis* CDC1551, this gene is fused to a hypothetical protein of unknown function owing to a mutation of $T_3 \Rightarrow T_2$ in hypothetical proteins. Further BLAST searches revealed that this fusion is not present in any other bacteria and hence seems specific to *M. tuberculosis* CDC1551.

Table 5.1: List of the fused genes from *M. tuberculosis* (H37Rv and CDC1551) and *M. bovis*. Fusions are shown in bold.

	M. tuberculosis H37Rv		*M. tuberculosis CDC1551*		*M. bovis*	
Genes	repeat	GI (mutation site, length of the ORF)	repeat	GI (mutation site, length of the ORF)	repeat	GI (mutation site, length of the ORF)
gmhA, Hypothetical protein (possible dehydratase)	T-5	15607255 (179, 196), 15607256 (190)	T-4	**15839494** (202, 420)	T-5	31791291 (179, 196), 31791292 (190)
fumarate reductase B *frdB*, fumarate reductase C *frdC*	G-11	15608691 (243, 247), 15608692 (126)	G-11	15841020 (243, 247), 15841021 (126)	G-7	**31792738** (243, 374)
Hypothetical protein, Hypothetical protein	G-5	15609698 (97), 15609699 (129)	G-6	**15842100** (212)	G-6	**31793744** (245)
Hypothetical protein, Hypothetical protein	G-6	15610016 (189), 15610017 (175, 275)	G-5	**15842421** (175, 364)	G-5	**31794056** (175, 364)
pks1, pks15	G-5	15610083 (1616), 15610084 (488, 496)	G-5	15842494 (1620), 15842495 (488, 496)	G-6	**31794123** (488, 2112)
Hypothetical protein, Hypothetical protein	CG-3	15610111 (470), 15610112 (83, 84)	CG-3	15842527 (470), 15842528 (91, 92)	CG-2	**31794151** (83, 553)
rfbB, Hypothetical protein	T-2	15610920 (326), 15610921 (325, 357)	T-3	**15843407** (325, 712)	T-2	31794957 (326), 31794958 (325, 357)

5.3.2 Gene fission

Of the 23 gene fission events observed, 20 were found in *M. bovis*, 2 were found in *M. tuberculosis* H37Rv and one was in *M. tuberculosis* CDC1551 (Table 5.2). Of the two fission events in *M. tuberculosis* H37Rv, one is found in the ORF encoding isocitrate lyase (*aceA*), which is an essential enzyme in the glyoxylate cycle that is required for growth with acetate and certain fatty acids as carbon sources (Bishai, 2000). The ORF *aceA* in *M. bovis* and *M. tuberculosis* CDC1551 harbors the T_5 tract which has undergone deletion of a T in *M. tuberculosis* H37Rv due to which there is premature termination at the 367^{th} codon. Because of this, the ORF *aceA* is split into two ORFs, (*aceAa* and *aceAb*), encoding isocitrate lyase alpha module (GI: 15609052) and isocitrate lyase beta module (GI: 15609053) (Cole et al. 1998). It is interesting to note that aceA forms one of the essential proteins and it remains to be tested whether the two split ORFs are functionally active or not. Nonetheless, all the three genomes have paralogs of isocitrate lyase in them which are probably acting as backup copies.

The other ORF that has undergone fission in *M. tuberculosis* H37Rv is a hypothetical protein. Equivalent ORF (un-split ORF) in *M. bovis* has been annotated as a probable conserved transmembrane protein and shown to contain a codB domain. CodB domain is seen in the *codB* gene which codes for cytosine transporter. Apart from this no other cytosine transporter has been found in the mycobacterial genomes (Cole et al. 1998). Cytosine is structurally related to one of the anti-mycobacterial drugs, pyrazinamide (Raynaud et al. 1999) and uptake system of pyrazinamide has been shown to uptake cytosine into mycobacterial cell (Raynaud et al. 1999). Hence, it can be surmised that ORF fission in the codB domain might affect uptake of pyrazinamide thereby leading to pyrazinamide resistance in mycobacteria.

In *M. bovis*, two of the polyketide synthases (*pks1* and *pks5*) are fused together and gaven rise to a single ORF (GI:31794123). This fusion is because of the insertion in the repeat tract ($G_5 \Rightarrow G_6$) in *M. tuberculosis* species. Polyke-

tide synthases are a class of enzymes, involved in the biosynthesis of secondary metabolites (polyketides) and complex lipids. There are a number of polyketide synthase (pks) genes found in the mycobacterial genomes (Cole et al. 1998). These genomes are known for their wide variety of polyketides (Cole et al. 1998). These polyketide synthases have also been shown to catalyze the last condensation step of mycolic acid biosynthesis (Damien Portevin et al. 2004). Alteration in this gene may cause changes to lipid and polyketide synthesis (Waddell et al. 2005).

As mentioned earlier *M. bovis* has harbored 20 instances of ORF fissions. One of the ORFs has been annotated as *fusA*2 and this encodes for translation elongation factor G (EF-G)- one of the key enzymes in protein synthesis. In this ORF fission has occurred due to repeat deletion ($G_3 \Rightarrow G_2$) at a region which codes for GTP_EFTU, an essential domain for GTP binding and thus probably making it non-functional. It has been observed that the organisms having mutations in the *fusA* gene are resistant to fusidic acid, a steroid antibiotic (Fuursted et al. 1992) which binds to the EF-G and inhibits the protein synthesis in the translocation of peptidyl tRNA from the A-site to the P-site (Tanaka et al. 1968). It has also been shown that minimum inhibitory concentration (MIC) and minimal bactericidal concentrations (MBC) of fusidic acid for *M. bovis* are twice that of *M. tuberculosis* (Fuursted et al. 1992) and this could be due to the non-functional gene *fusA2*.

In addition to ORF fusions, pks are also exhibiting ORF fissions. In *pks6*, fission is seen after 460^{th} amino acid that is in the end of pks domain because of an insertion ($G_2 \Rightarrow G_3$). Involvement of the *pks6* genes in virulence has been demonstrated previously (Camacho et al. 1999). Inactivation of this gene attenuates *M. tuberculosis* in mice (Camacho et al. 1999). The transport of pks is carried out by *mmpLS* gene products (Brosch et al. 2002). In *M. bovis*, fission is also seen in the middle of *mmpL*, an integral membrane protein, involved in the transport of polyketides as well as lipids (Garnier et al. 2003). Alteration in

these genes might cause changes to lipid and polyketide transport.

Other affected transport proteins are pstB, ugpA and a drug transporter (annotated as a hypothetical protein in *M. tuberculosis* H37Rv (GI:15609014)). pstB is a peripheral membrane protein that is involved in the uptake of phosphate, which is a part of a phosphate-specific transport (pst) operon (Cox et al. 1981). In *M. bovis*, the split occurs at the 64^{th} amino acid, owing to an insertion in poly T tract ($T_6 \Rightarrow T_7$). ugpA is an ATP-binding cassette transporter for sn-glycerol-3-phosphate and certain glycerophosphodiesters and is part of a ugp operon. (Su et al. 1991). In this ORF, the termination occurs at a codon corresponding to 85^{th} amino acid. In these two ORFs, mutations occur before the actual functional domain, which may not completely abolish the function of these proteins but could bring about certain changes in their properties. In the third protein which is the drug transporter, the split is seen within the transporter domain (from DOMAN FISHING) and one can expect functional nullification.

Fission has also been observed in the surface antigens, like the PE_PGRS family of proteins (Brennan et al. 2001) and envelope proteins like lppO. These changes provide clear proof for antigenic variation among the three mycobacteria. Membrane proteins and surface antigens are the first proteins to interact with antibodies and other host proteins. Modifications in them are very essential for a pathogen to survive in the host.

In all the proteins which have undergone fission, the ORF split at the domain linkers result in two independent domain units which in turn can express independently. In some cases split occurs within a domain which results in a non-functional short translational product. Of the 23 split ORFs examined, 7 were found with fission at the linkers of the known domains and 13 were found within known domains.

Though, the interactions of independently expressing individual domains are known to restore the function of their unified multi-domain counterpart, those with fissions within domains can not restore their unified counterparts. However,

genomes often harbor duplicate copies which act as backups. We discovered that all the fission genes which have lost function due to split within domains have paralogs (duplicate genes). The list of the fission genes and their paralogs are given in Table 5.3. (supplementary data for detailed report on repeating motifs, mutations and their locations).

5.4 Discussion

Variation in structural genes, especially in membrane and surface proteins offers a particular selection advantage for a pathogen to enable its pathoadaptations (Moxon et al. 1994). Variation due to microsatellite mutation in genes like *mmpL1, pks6, lppO, ugpA* and *fadE22* are previously observed and have been shown to become pseudogenes in *M. bovis* because of their premature termination (Garnier et al. 2003). However, as revealed by our studies premature terminations are also caused by indels of repeat units in the microsatellites. However, the microsatellite mutations are reversible and therefore it is quite probable that some of the pseudo genes may restore their functions. Fusions of *frdB, frdC and pks1* and *pks15* have also been reported in *M. bovis* (Garnier et al. 2003). These fused proteins may serve as intermediates for the emergence of novel proteins from subsequent adaptations (Long, 2000).

A high divergence is found in the type of polyketides produced by mycobacterial organisms. Some of the pks products have been observed only in the host and not in the laboratory conditions (Garnier et al. 2003). In the host, the organism is in a constrained environment and therefore mutation rates are expected to be high. Microsatellites are known to cause variations in genes under stress conditions. Thus, microsatellite variation may also play an important role in addition to giving a selection advantage under stress conditions.

Table 5.2: List of the split genes from *M. tuberculosis* (H37Rv and CDC1551) and *M. bovis*

	M. tuberculosis H37Rv		*M. tuberculosis CDC1551*		*M. bovis*	
Genes	repeat	GI (mutation site, length of the ORF)	repeat	GI (mutation site, length of the ORF)	repeat	GI (mutation site, length of the ORF)
*fusA*2	G-3	15607262 (106, 714)	G-3	15839501 (106, 714)	G-2	31791298 (597), 31791299 (106, 117)
*mmpL*1	C-4	15607543 (566, 958)	C-4	15839785 (566, 958)	C-5	31791578 (367), 31791579 (566, 591)
*pks*6	G-2	15607546 (456, 1402)	G-2	15839791 (456, 1402)	G-3	31791582 (456, 460), 31791583 (946)
Hypothetical protein	C-7	15607838 (101, 203)	C-7	Not annotated as gene	C-6	31791882 (101, 109), 31791883 (77)
Hypothetical protein	G-4	15607880 (52, 175)	G-3	15840149 (79, 82), 15840150 (120)	G-4	31791926 (52, 175)
Hypothetical protein	T-3	15607925 (134, 566)	T-3	15840200 (134,566)	T-2	31791972 (134, 191), 31791973 (368)
pstB	T-6	15608073 (64, 276)	T-6	15840356 (64, 276)	T-7	31792121 (64, 71), 31792122 (213)
PE_PGRS(wag22)	G-8	15608897 (914)	G-8	Not annotated as gene	G-7	31792948 (820), 31792949 (84, 94)
Hypothetical protein (drug transporter)	C-2	15609014 (267, 687)	C-2	15841347 (267, 687)	C-3	31793067 (267, 511), 31793068 (404)
Zinc-binding dehydrogenase	A-3	15609032 (96, 384)	A-3	15841366 (96, 384)	A-2	31793087 (96, 107), 31793088 (234)

(continued)

Table 5.2: (continued...)

aceA	T-4	15609052 (295, 367), 15609053 (398)	T-5	50953770 (295, 766)	T-5	31793108 (295, 766)
cobL	T-2	15609209 (96, 390)	T-2	15841561 (96, 390)	T-3	31793255 (294), 31793256 (42, 62)
lppO	C-6	15609427 (39, 171)	C-6	15841781 (60, 192)	C-5	31793468 (39, 51), 31793469 (121)
IS1558, transposase	TG-2	15609561 (67, 333)	TG-2	15841943 (67, 333)	TG-1	31793603 (230), 31793604 (67, 97)
PE_PGRS	G-7	15609627 (431, 1660)	G-7	15842017 (436, 1665)	G-15	31793669 (1150), 31793670 (431, 509)
Hypothetical protein (transglutaminase family protein)	GC-3	15609703 (531, 1140)	GC-3	15842104 (531, 1156)	GC-2	31793748 (531, 533), 31793749 (597)
ugpA	T-4	15609972 (85, 303)	T-4	15842376 (85, 303)	T-3	31794011 (180), 31794012 (84, 123)
*fadE*22	C-2	15610198 (110, 721)	C-2	15842629 (110, 721)	C-3	31794239 (600), 31794240 (110, 114)
Conserved hypothetical protein (mesT)	C-4	15610312 (199, 318)	C-4	15842753 (220, 339)	C-3	31794353 (105), 31794354 (200, 208)
Conserved transmembrane protein	T-3	15610589 (107, 110), 15610590 (422)	T-2	15843049 (108, 562)	T-2	31794630 (107, 561)
P450 heme-thiolate protein	C-4	15610654 (196, 398)	C-4	15843128 (170, 372)	C-3	31794693 (193), 31794694 (197, 205)
Hypothetical protein	A-6	15610909 (52, 194)	A-6	15843396 (52, 194)	A-7	31794945 (114), 31794946 (52, 78)
Hypothetical protein (FtsK/SpoIIIE family)	C-2	15611030 (770, 1396)	C-2	15843525 (770, 1396)	C-3	31795067 (561), 31795068 (770, 833)

Table 5.3: Paralogs for the split genes in *M. tuberculosis* (H37Rv and CDC1551) and *M. bovis* (identity is taken comparing unsplit gene and paralog)

Gene name	GI of Fission genes	GI of Paralogs (%ID)
M. tuberculosis H37Rv		
aceA	15609052, 15609053	15607608 (38)
M. tuberculosis CDC1551		
Hypothetical protein	15840149, 15840150	15840164 (91)
M. bovis		
*fusA*2	31791298, 31791299	31791868 (31)
*mmpL*1	31791578, 31791579	31792742 (58), 31791628 (57), 31791860 (56), 31791689 (54), 31793523 (54)
pks6	31791582, 31791583	31793231 (44), 317941231 (43), 31792849 (38), 31792851(49), 31794108 (35)
Hypothetical protein	31791882, 31791883	-
Hypothetical protein	31791972, 31791973	-
pstB	31792121, 31792122	31792008 (45), 31793575 (29), 31794928 376 (28), 31791839 (29), 31794833 (29)
PE_PGRS(wag22)	31792948, 31792949	31791933 (71), 31791457 (73), 31791456 (69), 31791458 (71), 31792021 (64)
Hypothetical protein (drug transporter)	31793067, 31793068	31793517 (29), 31792443 (25), 31794023 (26), 31793640 (26)
Zinc-binding dehydrogenase	31793087, 31793088	31794265 (28), 31791949 (27), 31791340 (27), 31793439 (29), 31792716 (26)
cobL	31793255, 31793256	-
lppO	31793468, 31793469	-
IS1558, transposase	31793603, 31793604	31793355 (98)
PE_PGRS	31793669, 31793670	31794687 (56), 31792644 (60), 31794684 (57)
Hypothetical protein (transglutaminase family protein)	31793748, 31793749	-
ugpA	31794011, 31794012	31793223 (31), 31793499 (32), 31792429 (26)
*fadE*22	31794239, 31794240	31791449 (43), 31794750 (43)
Conserved hypothetical protein (mesT)	31794353, 31794354	-
P450 heme-thiolate protein	31794693, 31794694	31793445 (36), 31791966 (33), 31794721 (28), 31791954 (30), 31793447(29)
Hypothetical protein	31794945, 31794946	31791758 (45), 31792915 (29)
Hypothetical protein (FtsK/SpoIIIE family)	31795067, 31795068	31792972 (29)

Gene duplication is a common phenomenon in bacterial genomes and is essential for evolution of the organisms. These duplicated genes have shown to increase gene expression diversity and speed of evolvement as compared to single-copy genes (Gu et al. 2004). Apart from acting as back-up copies to the existing genes, they can also tolerate mutations due to reduced mutational constraints, thus giving rise to increased likelihood of functional diversification (Gu et al. 2004). These mutations certainly increase the genetic diversity and may also drive the evolution of divergent proteins.

Microsatellite variation can lead to fusion of adjacent genes on the same strand. There is a good probability for operon genes to fuse together. Since the proteins are from the same operon, the expression levels would be balanced in the fused genes.

In general, horizontal gene transfers are significantly less in pathogenic bacteria (Garcia-Vallve et al. 2003) and are especially absent in *M. tuberculosis* (Cole et al. 1998). As it is extremely essential for pathogens to maintain heterogenity in their populations for survival, they adapt to different mechanisms. On the basis of our data, we propose that microsatellite induced affects could form one of the mechanisms to maintain heterogeneity in the mycobacterial genomes. As the microsatellites are present throughout these genomes their mutations can, in principle, fine-tune any gene if necessary, as mutations are very rapid and reversible.

Although the three mycobacterial genomes studied are 99% identical to each other at the level of nucleotides (Fleischmann et al. 2002; Garnier et al. 2003), our studies reveal different mutational constrains acting on them. In addition, these variations convey different adaptation forces working on these mycobacterial organisms that are necessary for their survival.

To conclude, we would like to emphasize that a large number of fission events are observed in *M. bovis* as compared to *M. tuberculosis*. Whether this has any relationship to its wide host range is a matter of further investigations which falls

outside the scope of this thesis work.

5.5 Summary

Gene fusion and fission are the most important events in the functional evolution of the proteins. These events are the consequences of the different mutations in the genomes. Here, we show for the first time the occurrence of gene fusion and fission events owing to microsatellites variations. In our whole genome comparative sequence analysis of *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv and *M. bovis*, we found a total of thirty genes with gene fusion and fissions. Analysis of the ORFs indicate that some of the fission events have led to the formation of pseudogenes. All the gene fusions and fissions are due to the indels of repeat units in mononucleotide and dinucleotide repeats. Our results confirm to an extent that microsatellite variations play a crucial role in the pathoadaptations of the pathogens.

Chapter 6

Role of point mutations in the origin and evolution of microsatellites

---

Publications from this chapter

- **Sreenu, V.B.**, Kumar, P. and Nagarajaram, H. A. (2005) Evolution of microsatellites through point mutations in *Mycobacterium tuberculosis* and *Mycobacterium bovis* (manuscript under preparation).

6.1 Introduction

As mentioned before, the unique property of microsatellites is their length polymorphism, which occurs because of mutations in the form of insertion and deletion of repeat motifs. Mutations occur due to strand slippage during DNA replication (Levinson and Gutman, 1987) that in turn has been shown to depend on repeat tract length (Dechering et al. 1998; Rose and Falush, 1998). Microsatellite tracts with repeat number less than six are less prone to mutations, and the mutation rate increases as the microsatellite tract gets longer (Dechering et al. 1998; Rose and Falush, 1998; Strassmann et al. 1997; Weber, 1990; Zhu et al. 2000a). However, point mutations in the long repeat tracts which disrupt repeat iterations reduce slippage rates (Petes et al. 1997). The other factors that influence microsatellite mutations include the length of the repeating motif (Chakraborty et al. 1997) and the repeating sequence (Eckert and Yan, 2000). Previous models on microsatellite mutations hypothesized microsatellite tracts to be in equilibrium with point mutations thus making them less prone to slippage mutations (Kruglyak et al. 1998; Primmer and Ellegren, 1998). The point mutations were shown to vary with distance from the origin of replication (Sharp et al. 1989). However, according to Hudson et al. (2002), the rate of point mutations vary between regions close to the origin of replication and regions far from it, with no correlation observed between the distance from the origin of replication and the rate of point mutations. (Hudson et al. 2002).

Since slippage can occur only in microsatellite tracts of certain length, it is intriguing as to how microsatellites originate at a locus and expand to a length sufficient to allow slippage errors. It is quite possible that short microsatellites are generated by random mutations; like insertion, deletions and substitutions in the genome (Levinson and Gutman, 1987; Messier et al. 1996; Rose and Falush, 1998; Stephan and Cho, 1994) or certain combinations of codons can give rise to microsatellites (Field and Wills, 1998). For example, the sequence of

codons for arginine and glycine amino acid residues viz., CGG and GGG gives rise to the microsatellite (G)5. Apart from random insertions and combinations of codons, microsatellites can also originate due to insertion of nucleotide stretches of adjacent nucleotides (Zhu et al. 2000b).

As mentioned earlier, point mutations lead to birth and death of microsatellite tracts (Messier et al. 1996; Rose and Falush, 1998). It is therefore interesting to study the effects of point mutations on the microsatellites especially in the genomes where there is no slippage error repair system. We therefore compared the genomic sequences of the three mycobacterial genomes *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv and *M. bovis* for microsatellites harboring point mutations and the details of this study are reported in this chapter. As will be found point mutations play a decisive role in the origin, morphogenesis and evolution of microsatellites.

6.2 Methods

The complete genome sequences of *M. tuberculosis* (CDC1551 and H37Rv) and *M. bovis* were devided into segments of 1500bp with an overlap of 50bp. These segments from one genome were searched against the complete genome sequences of other two organisms using BLASTN (version 2.2.6) (Altschul et al. 1997). Repeat masking filter (-F F) and a 100bp dropoff value for final gapped alignment (-Z 100) were used along with the default BLASTN search parameters. The hits that gave rise to alignments with the 100% identity except a single point mutation in the microsatellites were selected for further analysis. Insertion/deletions of a single nucleotide at the end of mononucleotide tracts were not considered in this study, because these changes might also arise due to slippage mutation. Nature of the mutation (transition, transversion and INDEL) was assigned to these loci manually. The point mutations that were increasing the tract length were termed as *microsatelite genesis* (ex: AAATAA$\Rightarrow$ A_6) where as, point mutations that were

leading to change in the repeat motif (AA TTT⇒AAA TT) or repeat sequence (ATAT AA ⇒ AT $(A)_4$) were termed as *microsatellite morphogenesis*.

6.3 Results

Pair-wise comparisons of the three genomes enabled us to detect 1528, 1565 and 759 point mutations in the microsatellite tracts, within the genome pairs *M. tuberculosis* H37Rv - *M. bovis*, *M. tuberculosis* CDC1551 - *M. bovis*, and *M. tuberculosis* H37Rv - *M. tuberculosis* CDC1551 respectively (Table 6.1). Among these point mutations 86-88% of point mutations are located in the coding regions indicating no particular bias towards coding or non-coding regions (coding density of the mycobacterial genomes is also 88-90%). Of the point mutations about 55-65% are due to transitions; 33-39% are due to transversions and 2-5% are due to INDELS.

The consequences of point mutations in microsatellites are illustrated in Fig. 6.1. Due to point mutations, some sequences, which may be called as promicrosatellites change to perfect microsatellites or in some cases short microsatellites expand and such events are collectively referred to as microsatellite genesis. For example, promicrosatellite sequence of ACGGTGCGGT in *M. tuberculosis* (at location 5752bp in both the strains CDC1551 and H37Rv) is becoming a microsatellite of $(ACGGT)_2$ in *M. bovis* (at location 5752bp) because of a transition from G ⇒ A. Of all the point mutations 80% are involved in microsatellite genesis. Point mutations in some microsatellites also lead to complete change in the repeat tract. For example: the dinucleotide repeat $(GC)_2$AGTA changes to GC$(GTA)_2$. Such events are referred to as microsatellite morphogenesis.

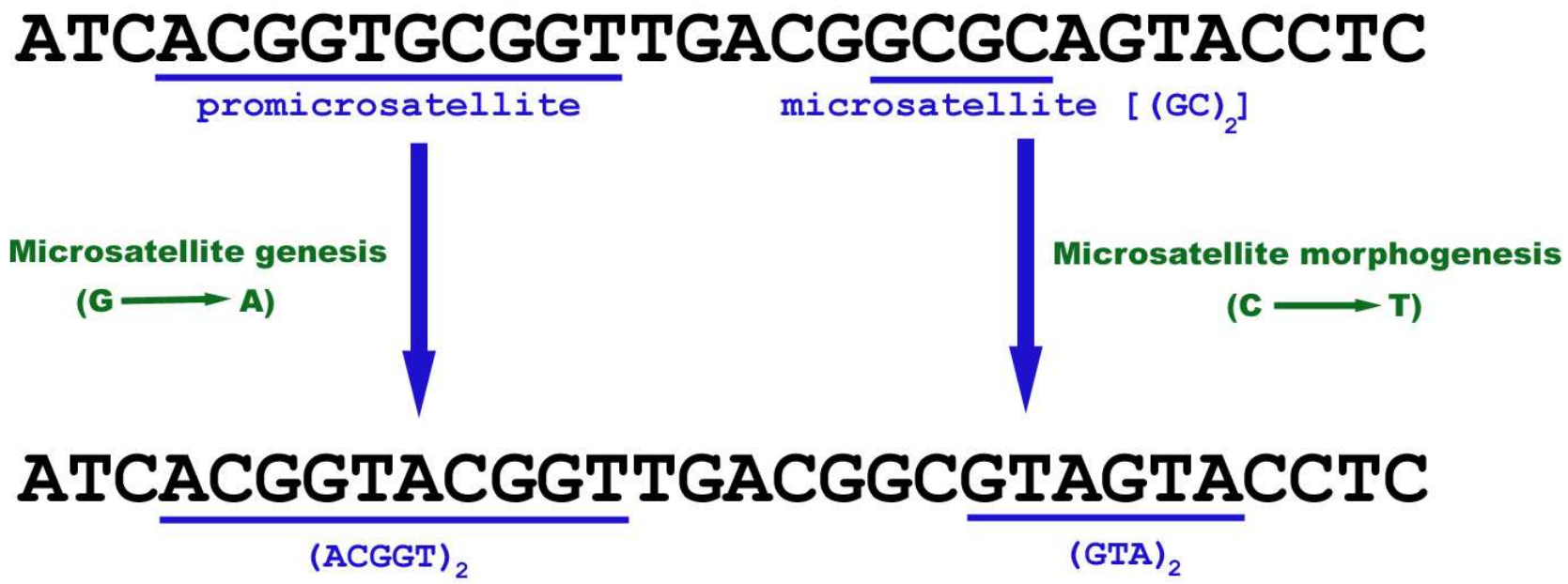

Figure 6.1: The consequences of point mutations in microsatellites

Between *M. tuberculosis* H37Rv and *M. bovis* 1197 microsatellite genesis and 331 examples of microsatellite morphogenesis events were found. Between *M. tuberculosis* CDC1551 and *M. bovis*, 1249 and 316 examples of microsatellite genesis and microsatellite morphogenesis respectively, were found. Between *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551 examples of 590 microsatellite genesis and examples of 169 microsatellite morphogenesis were found.

Nearly 25% of genes harbor the microsatellites with point mutations (supplementary data) of which about 50% are non-essential genes. The essential and non-essential gene definitions are as per Sassetti et al. (2003). Point mutations in most cases have led to the genesis of short microsatellite tracts. However, there are few loci where these mutations result in the genesis of long microsatellite tracts.

In the following some examples are given

The ORF *murG* in *M. bovis* contains a repeat of G_6 and the same gene in *M. tuberculosis* H37Rv as well as in *M. tuberculosis* CDC1551 contains $(G)_2A(G)_3$ indicating that this tract has been converted to a longer mononucleotide tract $(G)_6$ due to the transition of A$\Rightarrow$ G.

One of the hypothetical proteins of *M. bovis* contains the repeat G_8, which is found to be an imperfect tract (GGGGTGGG) in *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551. In MTC transglutamase, a tract of (T_7) is found as $CTC(T)_4$ in *M. tuberculosis* H37Rv and *M. bovis*. Two transitions in this tract have led to expansion of the repeat tract. In *M. bovis* repeat sequence C_6 is observed in one of the hypothetical proteins. The same is observed to be CCCCAC in *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551.

Table 6.1: Number of point mutations in microsatellites and their consequences among *M. tuberculosis* H37Rv (MTH), *M. tuberculosis* CDC1551 (MTC) and *M. bovis* (MB)

	Total (C:N)	Transitions	Transversions	Deletions
between *M. tuberculosis H37Rv* and *M. boivs*				
Total	1528 (1377:151)	995 (900:95)	499 (451:48)	34 (26:8)
MIC genesis	1197 (1081:116)	809 (732:77)	366 (331:35)	22 (18:4)
MIC morphogenesis	331 (296:35)	186 (168:18)	133 (120:13)	12 (8:4)
between *M. tuberculosis CDC1551* and *M. boivs*				
Total	1565 (1391:174)	987 (884:103)	521 (465:56)	57 (42:15)
MIC genesis	1249 (1102:147)	813 (725:88)	399 (353:46)	37 (24:13)
MIC morphogenesis	316 (289:27)	174 (159:15)	122 (112:10)	20 (18:2)
between *M. tuberculosis H37Rv* and *M. tuberculosis CDC1551*				
Total	759 (664:95)	423 (380:43)	296 (264:32)	40 (20:20)
MIC genesis	590 (519:71)	343 (310:33)	220 (197:23)	27 (12:15)
MIC morphogenesis	169 (145:24)	80 (70:10)	76 (67:9)	13 (8:5)

(The tracts are available as supplementary data)

6.4 Discussion

From our comparative genome analysis it is found that point mutations can create sufficiently long tracts which can further undergo slippage. Messier et al. (1996) earlier demonstrated that G⇒A transition, changes the sequence ATGTGTGT to the ATGTATGT $(ATGT)_2$ microsatellite which further evolved into $(ATGT)_4$ and $(ATGT)_5$ in descending lineages. And at the same loci, the transition of A⇒G brought about the change of sequence ATGTGTGT to GTGTGTGT $(GT)_5$ which subsequently evolved to become GT6 in further generations.

Several inter-conversions have been observed between tracts, like in place of dinucleotide repeat tracts, trinucleotide or mononucleotide repeat tracts appear. Instances where one microsatellite dies and the other expands have also been observed. All these changes have an important implication that the rates of slippage which give rise to microsatellite length polymorphisms, change because the slippage error rates depend on the sequence type and length as shown by Weber and Wong (1993), Chakraborthy et al. (1997) and (Eckert and Yan, 2000). Taking cue from these, it can be expected that point mutations that lead to motif changes in terms of size and sequence, give rise to changes in polymorphism rates and this can be one of the ways of brining plasticity to the genomes.

Morphogenesis of trinucleotide repeat to a dinucleotide or a tetranucleotide repeat, can have a major effect on the ORF function due to subsequent slippage mutations, that can in turn lead to frameshifts in the reading frame. These changes can trigger many ORFs to become "contingency loci", thus allowing the organism to make better use of them for its adaptation via the slippage mechanism. One of the best examples of this type of mutations is microsatellite morphogenesis from $GCC_3 \Rightarrow C_5$ in integral membrane protein. In*M. tuberculosis* CDC1551 and *M. bovis* one of the integral membrane proteins (GI 15839705 and 31791497 respectively) harbors a microsatellite GCCGCCGCC $[(GCC)_3]$ at location 386492bp and 387462bp respectively. In this locus, transversion from

G ⇒ C creates a mononucleotide tract of C_5 in *M. tuberculosis* H37Rv. Slipage mutations in this mononucleotide tract lead to frame-shifts in the reading frame, thus creating variants of integral membrane proteins.

Non-essential genes are very important for the organism's adaptation and survival. Their tolerance to mutations help fine tuning and shaping of the organism to thrive and adapt in adverse conditions. Our investigations have revealed that non-essential genes are gaining longer microsatellite tracts than essential genes (supplementary data) which could be a means of aiding the ORFs to face difficult environments through rapid changes via slippage. The major advantage with the slippage mutation is its reversible nature. An organism can therefore bring back its previous microsatellites in subsequent generations.

6.5 Summary

In this chapter we presented a comparative analysis of the three genomes *M. bovis*, *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551 for point mutations in the microsatellites and discussed the nature of point mutations and their effects on the microsatelites. Due to point mutations the genomes acquire microsatellites (microsatellite genesis) which can potentially become polymorphic. Point mutations also lead to microsatellite morphogenesis where one class of microsatellite is substituted by another class of microsatellite. This has an important implication on the genome in particular the coding regions nearby, as the rates of mutations in microsatellites are dependent on the size and sequence. Hence in addition to the microsatellites point mutations in them, they also impart additional plasticity to the pathogenic genomes.

Appendix A

PBD DNA Interactions

Published papers from this chapter

- A. Kamal, N. Laxman, G. Ramesh, P. Ramulu, O. Srinivas, K. Neelima, A. K. Kondapi, **V. B. Sreenu** and H.A.Nagarajaram(2002) Design, synthesis, and evaluation of new non-cross linking pyrrolobenzodiazepine dimers with efficient DNA-binding ability and potent antitumour activity *J. Med. Chem* 45:4679-4688.

A.1 Introduction

There has been considerable interest in the design and development of DNA adducts that are capable of binding to DNA in a sequence selective manner. However, in spite of rational design of synthetic DNA intercalaters, only few such compounds that are in current clinical use exhibit sequence selectivity. These compounds with the ability to target and then down regulate individual genes have potential use as drugs for therapy of genetic-based diseases including some cancers and as research tools for using functional genomic studies. Therefore, the synthesis of small molecules, which exhibit DNA sequence selectivity, is of importance for the targeting of rapidly growing tumor cells.

The pyrrolo[2,1-c][1,4]benzodiazepine (PBD) antitumor antibiotics are a well-known class of sequence selective DNA binding agents derived from *Streptomyces* species. Their interactions with DNA are unique since they bind within the minor groove of DNA forming a covalent aminal bond between the C11 position of the central B-ring and the N2 amino group of a guanine base. The cytotoxic and antitumor activity of PBDs are attributed to their ability to form covalent DNA adducts. The PBDs have been shown to inhibit endonuclease enzyme cleavage of DNA and to block transcription by inhibiting DNA polymerase in a sequence specific manner the processes which may be relevant for the biological activity.

A.2 Methods

To Study the DNA and PDB intercations we carried out molecular simulation studies. All of the calculations were performed using INSIGHT-II suite of software (MSI, Inc.) running on a Silicon Graphics OCTANE system.

A.2.1 Modeling of DNA Duplex and PBD Dimer Structures

The 15-mer sequence GGGGCGAGAGAGGGG with the central AGA triplet representing the preferred binding site Pu- G-Pu for the PBD molecule was chosen for modeling the B-DNA duplex structure. The model was constructed using the BIOPOLYMER module (Fig. A.1). Models for the PBD dimers 5a-d were built using the BUILDER module. First, the PBD molecules were 'sketched' in two dimensions (2D) and then converted into three dimensional (3D) entities using the 2D to 3D conversion tool in the BUILDER module. During modeling, it was assumed that both of the PBD units are having (S)-stereochemistry at the chiral C11 atom.

A.2.2 Docking Studies

Docking of a PBD dimer into the minor groove of the DNA duplex was carried out manually such that the N10-C11 imine functionality and the exocyclic C2-amino group of G8 are nearly at a bonding distance from each other so that a covalent bond can be formed. After the covalent bond was created, the PBD dimer in the minor groove was manually oriented in such a way that the PBD with amide functionality is oriented toward the 3 end of a covalently linked DNA strand. Some of the dihedral angles about the C-C bonds of the spacer units were manually adjusted such that the entire PBD dimer has an isohelical fit within the minor groove of the DNA duplex. The potentials and the charges were fixed using the CVFF force field. The complex so formed was subjected to energy minimization using the conjugate gradient technique until full convergence (energy gradient e 0.001) was reached. During energy minimization and MD simulations, constraints were applied to fix the DNA duplex structure.

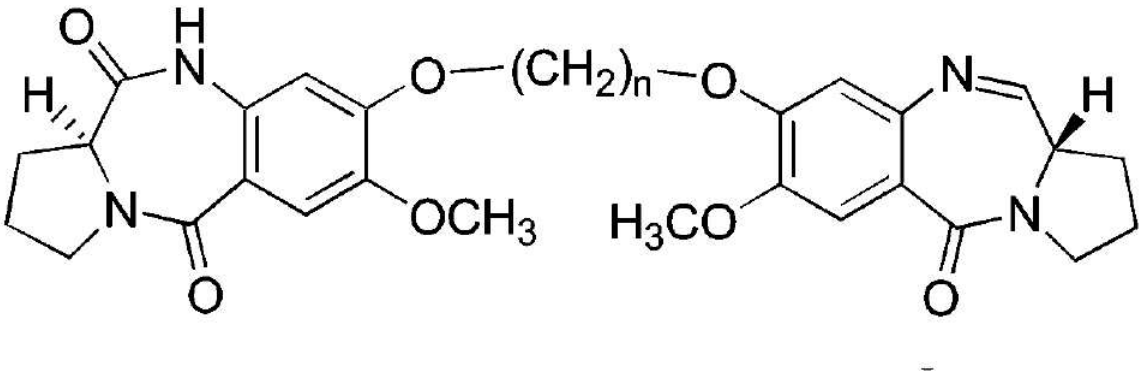

Figure A.1: Chemical structure of PBD molecule with alkane spacer units of number n=3(5a), 4(5b), 5(5c) and 8(5d)

A.2.3 Molecular Dynamics Simulations

The following protocol was used to carry out the molecular dynamics studies: heating phase (equilibration)= 0-10 ps and length of MD simulation after the heating phase = 100 ps. During simulations, all the intermittent structures of the complexes formed at every successive 5 ps were saved and later they were subjected to energy minimization. The minima obtained for each complex were screened to identify the lowest energy minimum and that was taken as a representative of energetically favorable complex for further studies. Energy of interaction between the DNA and the PBD dimer molecule in a complex was calculated as follows:

$E_{int} = E_{complex} - (E_{DNA} + E_{PBD})$

where E_{int}= energy of interaction of a complex, $E_{complex}$ = total energy of the complex, and (E_{DNA} and E_{PBD} are the individual total energies of the DNA and the PBD dimer molecules calculated after they are separated from each other.

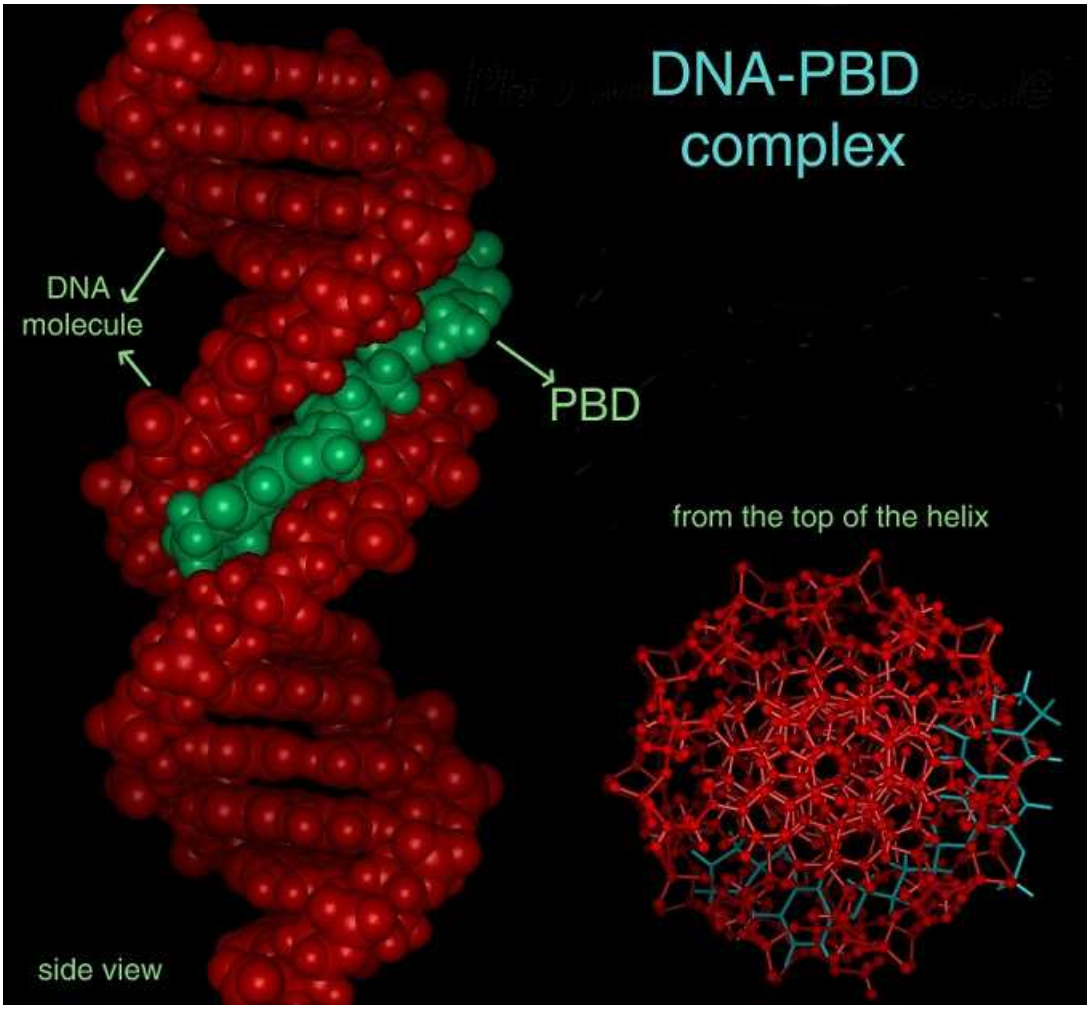

Figure A.2: DNA-PBD interaction

Table A.1: Values of Energy of Interactions Calculated for the DNA–PBD Dimer Complexes

complex	energy of interaction (E_{int}) in kcal mol^{-1}
DNA–**5a**	−123.4
DNA–**5b**	−129.2
DNA–**5c**	−139.7
DNA–**5d**	−137.5

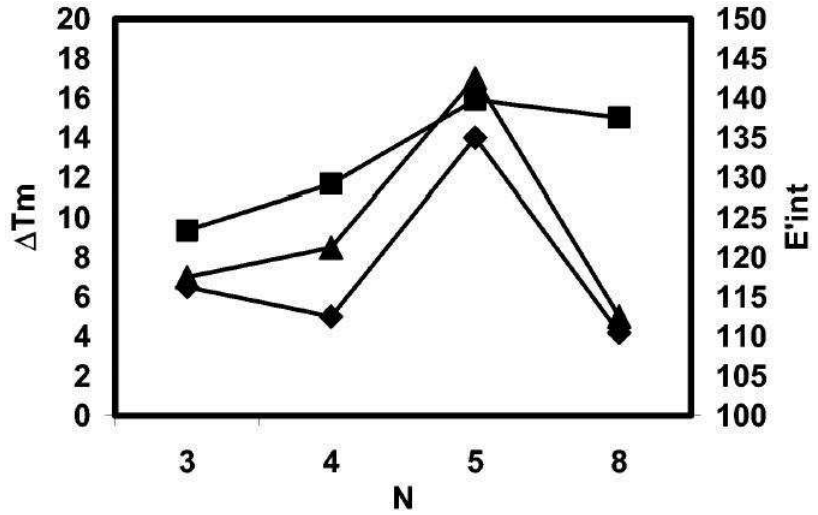

Figure A.3: Plot of the calculated values of energy of interactions (E_{int} in kcal mol^{-1}, ■) and the melting temperatures ($\triangle T_{\mathrm{m}}$ in °C, ◆ (0 h) and ▲ (18 h)) measured from thermal denaturation studies for the four DNA−PBD dimer complexes vs the number of spacer units.

A.3 Results and Discussion

A.3.1 Molecular Modeling Studies

Energetically favorable models of the DNA-PBD-dimer complexes for 5a-d molecules were built using the systematic procedure as described in the methods section, involving molecular modeling and docking followed by detailed molecular dynamic simulations. The energy of interaction (E_{int}) between the DNA and the PBD-dimer molecule in a complex was calculated as a measure of stability of that complex. Table below gives the values of E_{int} obtained for the four complexes. The lowest value of E_{int} corresponds to the DNA-5c complex indicating that the molecule 5c forms energetically the most stable complex with DNA (Fig A.2). It is interesting to note that the E_{int} progressively decreases indicating increase in stability as the number of alkane spacer units increases from three to five. However, further increase in the number of spacer units from five to eight does not lead to an increase in the E_{int}; in fact, E_{int} is lowered by about 2 kcal mol^{-1}. The E_{int} values are, in fact, correlated to the Tm values measured from thermal denaturation experiments (Table A.1 and Fig. A.3). In general, a number of favorable van der Waals and Coulombic interactions are formed between the DNA and the PBD-dimer molecules and they together contribute to the stability in all of the complexes, in addition to the covalent linkage formed between the imine PBD subunit and the G8. Only in the cases of 5c,d, the complexes are further stabilized by a hydrogen bond formed between the carbonyl oxygen of the C10-N11 amide functionality of the PBD subunit and the amino group of the 12^{th} guanine nucleotide. It is worth noting that the amide PBD subunits in both the complexes of 5c,d occupy the same site in the minor groove. The complex formed by 5c is energetically more stable than the complex formed by 5d because the chain of alkane spacer units in 5c forms a better isohelical fit (and thus gives rise to more favorable nonbonded interactions) within the minor groove than the longer chain in 5d.

A.4 Conclusion

From the systematic conformational searches involving molecular dynamics simulations we were able to obtain energetically favorable complex structures for DNA with a number of PBD dimmers each differing in the number of alkane spacer units. The models satisfactorily accounted for the observed relative stabilities of the different DNA-PDB complexes as demonstrated by correlation between their DNA melting temperatures and binding interaction energies of DNA and PBD molecules. The models were further analyzed to understand the underlying molecular interactions contributing to the relative stabilities of different DNA-PBDs. Lessons learnt in this study enabled our collaborators to design yet another set of novel PBD molecules.

References

1. Aaltonen, L.A., Peltomaki, P., Leach, F.S., Sistonen, P., Pylkkanen, L., Mecklin, J.P., Jarvinen, H., Powell, S.M., Jen, J., and Hamilton, S.R. (1993) Clues to the pathogenesis of familial colorectal cancer. *Science* **260:** 812-816.

2. Acharya, S., Foster, P.L., Brooks, P., and Fishel, R. (2003) The Coordinated Functions of the E. coli MutS and MutL Proteins in Mismatch Repair. *Mol. Cell* **12:** 233-246.

3. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

4. Amos, W., and Rubinstzein, D.C. (1996) Microsatellites are subject to directional evolution. *Nat. Genet.* **12:** 13-14.

5. Andrewes, F.W. (1922) Studies in group-agglutination. I. The Salmonella group and its antigenic structure. *J. Pathol. Bacteriol.* **25:** 505-521.

6. Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., and Tsilfidis, C. (1992) Cloning of the essential myotonic dystrophy region and mapping of the putative defect. *Nature* **355:** 548-551.

7. Bachtrog, D., Weiss, S., Zangerl, B., Brem, G., and Schlotterer, C. (1999) Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. *Mol. Biol. Evol.* **16:** 602610.

8. Banu, S., Honore, N., Saint-Joanis, B., Philpott, D., Prevost, M.C., and Cole, S.T. (2002) Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens? Mol. *Microbiol.* **44:** 9-19.

9. Bauer, B.A., Stevens, M.K., and Hansen, E.J. (1998) Involvement of the Haemophilus ducreyi gmhA Gene Product in Lipooligosaccharide Expression and Virulence. *Infect. Immun.* **66:** 4290-4298.

10. Bayliss, C.D., van de Ven, T., and Moxon, E.R. (2002) Mutations in polI but not mutSLH destabilize Haemophilus influenzae tetranucleotide repeats. *EMBO J.* **21:** 1465-1476.

11. Bell, G.I., and Jurka, J. (1997) The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *J. Mol. Evol.* **44:** 414421.

12. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

13. Bishai, W. (2000) Microbiology: Lipid lunch for persistent pathogen. *Nature* **406:** 683-685.

14. Bizzaro, J. W. and Marx, K. A. (2003) Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *BMC Bioinfo.* **4:** 22.

15. Blanchard, J.S. (1996) Molecular mechanisms of drug resistance in Mycobacterium tuberculosis. *Annu. Rev. Biochem.* **65:** 215-239.

16. Brais, B., Bouchard, J.P., and Xie, Y.G. (1998) Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* **18:** 164-167.

17. Brennan, M.J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M., and Jacobs Jr, W.R. (2001) Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that Influence interactions with other cells. *Infect. Immun.* **69:** 7326-7333.

18. Breslauer, K.J., Frank, R., Blocker, H., and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. U. S. A.* **83:** 3746-3750.

19. Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., and Rolf, B. (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62:** 1408-1415.

20. Brooker, A.L., Cook, D., Bentzen, P., Wright, J.M., and Doyle, R.W. (1994) Organisation of microsatellites differs between mammals and cold-water teleost fishes. *Can. J. of Fish. Aqua. Sci.* **51:** 1959-1966.

21. Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K. et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 3684-3689.

22. Buchanan, G., Irvine, R.A., Coetzee, G.A., and Tilley, W.D. (2001) Contribution of the androgen receptor to prostate cancer predisposition and progression. *Cancer Metastasis Rev.* **20:** 207-233.

23. Burch, C.L., Danaher, R.J., and Stein, D.C. (1997) Antigenic variation in Neisseria gonorrhoeae: production of multiple lipooligosaccharides. *J. Bacteriol.* **179:** 982-986.

24. Burge, C., Campbell, A.M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci U S A* **89:** 1358-1362.

25. Calabrese, P.P., Durrett, R.T., and Aquadro, C.F. (2001) Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* **159:** 839852.

26. Calmann, M.A., Evans, J.E., and Marinus, M.G. (2005) MutS inhibits RecA-mediated strand transfer with methylated DNA substrates. *Nucleic Acids Res.* **33:** 3591-3597.

27. Camacho, L.R., Ensergueix, D., Perez, E., Gicquel, B., and Guilhot, C. (1999) Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. *Mol. Microbiol.* **34:** 257-267.

28. Campuzano, V., Montermini, L., Molto, M.D., Pianese, L., Cossee, M., and Cavalcanti, F. (1996) Friedreichs ataxia:autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271:** 1423-1427.

29. Castelo, A. T., Martins, W., and Gao, G. R. (2002) TROLLTandem Repeat Occurrence Locator. *Bioinformatics* **18:** 634-636.

30. Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., and Deka, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. U.S.A.* **94:** 1041-1046.

31. Chambers, A.L., Smith, A.J., and Savery, N.J. (2003) A DNA translocation motif in the bacterial transcriptionrepair coupling factor, Mfd. *Nucleic Acids Res.* **31:** 6409-6418.

32. Chambers, G.K., and MacAvoy, E.S. (2000) Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. B.* **126:** 455-476.

33. Cifarelli, R. A., Gallitelli, M., Cellini, F. (1995) Random amplified hybridization microsatellites (RAHM): isolation of a new class of microsatellite-containing DNA clones. *Nucleic Acids Res* **23:** 3802 3803.

34. Chubb, A.J., Woodman, Z.L., da Silva Tatley, F.M., Hoffmann, H.J., Scholle, R.R., and Ehlers, M.R. (1998) Identification of Mycobacterium tuberculosis signal sequences that direct the export of a leaderless beta-lactamase gene product in Escherichia coli. *Microbiology* **144:** 1619-1629.

35. Coenye, T., and Vandamme, P. (2003) Simple sequence repeats and compositional bias in the bipartite Ralstonia solanacearum GMI1000 genome. *BMC Genomics* **4:** 10.

36. Coetzee, G., and Irvine, R. (2002) Size of the androgen receptor CAG repeat and prostate cancer: does it matter? J. *Clin. Oncol.* **20:** 3572-3573.

37. Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.et al (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537-544.

38. Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D. et al (2001) Massive gene decay in the leprosy bacillus. *Nature* **409:** 1007-1011.

39. consortium, I.h.g.s. (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

40. consortium, M.g.s. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

41. consortium, R.g.s.p. (2004) Genome sequencing of the brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.

42. Contreras-Moreira, B., and Bates, P.A. (2002) Domain Fishing: a first step in protein comparative modelling. *Bioinformatics* **18:** 1141-1142.

43. Coreno, G., Ginelli, E., and E, P. (1967) A satellite DNA band isolated from human tissue. *J. Mol. Biol.* **23:** 619-622.

44. Cosma, C.L., Sherman, D.R., and Ramakrishnan, L. (2003) The secret lives of the pathogenic mycobacteria. *Annu Rev Microbiol* **57:** 641-676.

45. Cox, G.B., Rosenberg, H., Downie, J.A., and Silver, S. (1981) Genetic analysis of mutants affected in the Pst inorganic phosphate transport system. *J. Bacteriol.* **148:** 1-9.

46. Cummings, C.J., and Zoghbi, H.Y. (2000) Trinucleotide repeats:mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet* **1:** 281-328.

47. Damien-Portevin, D., de Sousa-D'Auria, C., Houssin, C., Grimaldi, C., Chami, M., Daff, M., and Guilhot, C. (2004) A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc. Natl. Acad. Sci. USA* **101:** 314-319.

48. Dechering, K.J., Cuelenaere, K., Konings, R.N., and Leunissen, J.A. (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res* **26:** 4056-4062.

49. Dejager, S., Bry-Gauillard, H., Bruckert, E., Eymard, B., Salachas, F., LeGuern, E., Tardieu, S., Chadarevian, R., Giral, P., and Turpin, G. (2002) A comprehensive endocrine description of Kennedys disease revealing androgen insensitivity linked to CAG repeat length. *J. Clin. Endocrinol. Metab.* **87:** 893-901.

50. de la Cruz, F., and Davies, J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8:** 128-133.

51. Di Rienzo, A., Peterson, A. C., Gakza, J. C., Vldes, A. M., and Slatkin, M. (1994) Mutational processes of simplesequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* **91:** 31663170.

52. Dong, C., Beis, K., Giraud, M.F., Blankenfeldt, W., Allard, S., Major, L.L., Kerr, I.D., Whitfield, C., and Naismith, J.H. (2003) A structural perspective on the enzymes that convert dTDP-d-glucose into dTDP-l-rhamnose. *Biochem. Soc. Trans.* **31:** 532-536.

53. Duval, A., and Hamelin, R. (2002) Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. *Cancer Res.* **62:** 2447-2454.

54. Eckert, K.A., and Yan, G. (2000) Mutational analyses of dinucleotide and tetranucleotide microsatellites in Escherichia coli: influence of sequence on expansion mutagenesis. *Nucleic Acids Res* **28:** 2831-2838.

55. Ejima, Y., Yang, L., and Sasaki, M.S. (2000) Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumor cell lines: a novel mutation target of microsatellite instability. *Int. J. Cancer* **86:** 262-268.

56. Ellegren, H. (2000) Microsatellite mutations in the germline:implications for evolutionary inference. *Trends Genet.* **16:** 551558.

57. Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nature Rev. Genet.* **5:** 435-445.

58. Enright, A.J., Iliopoulos, I., Kyrpides, N., and Ouzounis, C.A. (1999) Protein integration maps for complete genomes based on gene fusion events. *Nature* **402:** 86-90.

59. Feldman, M.W., Bergman, A., Pollock, D.D., and Goldstein, D.B. (1997) Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145:** 207216.

60. Ferro, P., dellEva, R., and Pfeffer, U. (2001) Are there CAG repeat expansion-related disorders outside the central nervous system? *Brain Res. Bull.* **56:** 259-264.

61. Field, D., and Wills, C. (1998) Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci U S A* **95:** 1647-1652.

62. Fisher, P. J., Gardner, R. C., Richardson, T. E. (1996) Single locus microsatellites isolated using 5'-anchored PCR. *Nucleic Acids Res* **24:** 4369 4371.

63. Flche, P.L., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., and Vergnaud, G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *BMC Microbiol.* **1:** 2.

64. Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., et al. (2002) Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *J Bacteriol* **184:** 5479-5490.

65. Foster, P.L., and Trimarchi, J.M. (1994) Adaptive reversion of a frameshift mutation in Escherichia coli by simple base deletions in homopolymeric runs. *Science* **265:** 407-409.

66. Franke, P., Leboyer, M., and Gansicke, M. (1998) Genotype-phenotype relationship in female carriers of the premutation and full mutation of FMR-1. *Psychiatry Res.* **80:** 113-127.

67. Fresco, J.R., and Alberts, B.M. (1960) The accommodation of noncomplementary bases in helical polyribonucleotides and deoxyribonucleic acids. *Proc. Natl. Acad. Sci. USA.* **46:** 311-321.

68. Fuursted, K., Askgaard, D., and Faber, V. (1992) Susceptibility of strains of the Mycobacterium tuberculosis complex to fusidic acid. *APMIS* **100:** 663-667.

69. Fu, Y.H., Kuhl, D.P., Pizzuti, A., Pieretti, M., Sutcliffe, J.S., and Richards, S. (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67:** 1047-1058.

70. Galas, D.J., Eggert, M., and Waterman, M.S. (1985) Rigorous pattern recognition methods for DNA sequences. *J. Mol. Biol* **186:** 117-128.

71. Galio, L., Bouquet, C., and Brooks, P. (1999) ATP hydrolysis-dependent formation of a dynamic ternary nucleoprotein complex with MutS and MutL. *Nucleic Acids Res.* **27:** 2325-2331.

72. Garcia-Vallve, S., Guzman, E., Montero, M.A., and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.* **31:** 187-189.

73. Garcia-Vallve, S., Romeu, A., and Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10:** 1719-1725.

74. Garnier, T., Eiglmeier, K., Camus, J.C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., et al. (2003) The complete genome sequence of Mycobacterium bovis. *Proc Natl Acad Sci U S A* **100:** 7877-7882.

75. Garza, J.C., Slatkin, M., and Freimer, N.B. (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* **12:** 594603.

76. Ge, Z., Jiang, Q., Kalisiak, M.S., and Taylor, D.E. (1997) Cloning and functional characterization of Helicobacter pylori fumarate reductase operon comprising three structural genes coding for subunits C, A and B. *Gene* **204:** 227-234.

77. Glickman, M.S., Cahill, S.M., and Jacobs Jr, W.R. (2001) The Mycobacterium tuberculosis cmaA2 gene encodes a mycolic acid trans-cyclopropane synthetase. *J. Biol. Chem.* **276:** 2228-2233.

78. Gordon, A.J., and Halliday, J.A. (1995) Inversions with deletions and duplications. *Genetics* **140:** 411-414.

79. Grimwood, J., Olinger, L., and Stephens, R.S. (2001) Expression of Chlamydia pneumoniae polymorphic membrane protein family genes. *Infect. Immun.* **69:** 2383-2389.

80. Groathouse, N.A., Rivoire, B., Kim, H., Lee, H., Cho, S.N., Brennan, P.J., and Vissa, V.D. (2004) Multiple polymorphic loci for molecular typing of strains of Mycobacterium leprae. *J Clin Microbiol* **42:** 1666-1672.

81. Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M., and Kashi, Y. (2000) Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. *Genome Res* **10:** 62-71.

82. Gu, Z., Rifkin, S.A., White, K.P., and Li, W.H. (2004) Duplicate genes increase gene expression diversity within and between species. *Nat. Genet.* **36:** 577-579.

83. Hallet (2001) Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr. Opin. Microbiol.* **4:** 570-581.

84. Halling, K.C., Harper, J., Moskaluk, C.A., Thibodeau, S.N., Petroni, G.R., Yustein, A.S., Tosi, P., Minacci, C., Roviello, F., Piva, P., et al. (1999) Origin of microsatellite instability in gastric cancer. *Am. J. Pathol.* **155:** 205-211.

85. Hancock, J.M. (1996) Simple sequences in a "minimal" genome. *Nat. Genet.* **13:** 14-15.

86. Harr, B., Zangerl, B., Brem, G., and Schlotterer, C. (1998) Conservation of locus-specific microsatellite variability across species: a comparison of two Drosophila sibling species, D. melanogaster and D. simulans. *Mol. Biol. Evol.* **15:** 176184.

87. Hauge, X.Y., and Litt, M. (1993) A study of the origin of shadow bands seen when typing dinucleotide repeat polymorphisms by the PCR. *Hum. Mol. Genet.* **2:** 411-415.

88. Haydon, A.M., and Jass, J.R. (2002) Emerging pathways in colorectal-cancer development. *Lancet Oncol.* **3:** 83-88.

89. Hermans, P.W.M., van Soolingen, D., Bik, M., de Haas, P.E.W., Dale, J.W., and van Embden, J.D.A. (1991) The insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in M. tuberculosis complex strains. *Infect Immun* **59:** 2695-2705.

90. High, N.J., Deadman, M.E., and Moxon, E.R. (1993) The role of the repetitive DNA motif (5'-CAAT-3') in the variable expression of the H. influenzae lipopolysaccharide epitope alpha-Gal(1-4)beta-Gal. *Mol. Microbiol.* **9:** 1275-1282.

91. Hile, S.E., and Eckert, K.A. (2004.) Positive correlation between DNA polymerase -primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *J. Mol. Biol* **335:** 745-759.

92. Hinrichs, W., Kisker, C., Duvel, M., Muller, A., Tovar, K., Hillen, W., and Saenger, W. (1994) Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. *Science* **264:** 418-420.

93. Honer Zu Bentrup, K., Miczak, A., Swenson, D.L., and Russell, D.G. (1999) Characterization of activity and expression of isocitrate lyase in Mycobacterium avium and Mycobacterium tuberculosis. *J. Bacteriol.* **181:** 7161-7167.

94. Hood, D.W., Deadman, M.E., Jennings, M.P., Bisercic, M., Fleischmann, R.D., Venter, J.C., and Moxon, E.R. (1996) DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc Natl Acad Sci U S A* **93:** 11121-11125.

95. Hudson, R.E., Bergthorsson, U., Roth, J.R., and Ochman, H. (2002) Effect of chromosome location on bacterial mutation rates. *Mol. Biol. Evol.* **19:** 85-92.

96. Inzana, T.J., Hensley, J., McQuiston, J., Lesse, A.J., Campagnari, A.A., Boyle, S.M., and Apicella, M.A. (1997) Phase variation and conservation of lipooligosaccharide epitopes in Haemophilus somnus. *Infect. Immun.* **65:** 4675-4681.

97. Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993) Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363:** 558-561.

98. Jardine, O., Gough, J., Chothia, C., and Teichmann, S.A. (2002) Comparison of the small molecule metabolic enzymes of Escherichia coli and Saccharomyces cerevisiae. *Genome Res.* **12:** 916-929.

99. Jarne, P., and Lagoda, P.J.L. (1996) Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.* **11:** 424-429.

100. Jeffreys, A.J., Wilson, V., and Thein, S.L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* **314:** 67-73.

101. Kamebeek, J. L. S., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., and van Embden, J.D.A. (1997) Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol* **35:** 907-914.

102. Kampke, T., Kieninger, M., and Mecklenburg, M. (2001) Efficient primer design algorithms. *Bioinformatics* **17:** 214-225.

103. Kant, J.A., Fornace, A.J.J., Saxe, D., Simon, M.I., McBride, O.W., and Crabtree, G.R. (1985) Evolution and organization of the fibrinogen locus on chromosome 4: gene duplication accompanied by transposition and inversion. *Proc. Natl. Acad. Sci. U. S. A.* **82:** 2344-2348.

104. Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283-290.

105. Karlin, S., Mrazek, J., and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179:** 3899-3913.

106. Katti, M.V., Ranjekar, P.K., and Gupta, V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol* **18:** 1161-1167.

107. Kenneson, A., Zhang, F., Hagedorn, C.H., and Warren, S.T. (2001) Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate length and premutation carriers. *Hum. Mol. Genet.* **10:** 1449-1454.

108. Kimmel, M., and Chakraborty, R. (1996) Measures of variation at DNA repeat loci under a general stepwise mutation model. *Theor. Popul. Biol.* **50:** 345367.

109. Kimmel, M., Chakraborty, R., Stivers, D.N., and Deka, R. (1996) Dynamics of repeat polymorphisms under a forwardbackward mutation model: within- and between-population variability at microsatellite loci. *Genetics* **143:** 549555.

110. Kisker, C., Hinrichs, W., Tovar, K., Hillen, W., and Saenger, W. (1995) The complex formed between Tet repressor and tetracycline-Mg2+ reveals mechanism of antibiotic resistance. *J. Mol. Biol* **247:** 260-280.

111. Klement, I.A., Skinner, P.J., Kaytor, M.D., Yi, H., Hersch, S.M., and Clark, H.B. (1998) Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. *Cell* **95:** 41-53.

112. Kornberg, A., Bertsch, L.L., Jackson, J.F., and Khorana, H.G. (1964) Enzymatic synthesis of deoxyribonucleic acid. XVI. Oligonucleotides as templates and the mechanisms of their replication. *Proc. Natl. Acad. Sci. USA.* **51:** 315-323.

113. Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., and Baker, E. (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* **252:** 1711-1714.

114. Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95:** 10774-10778.

115. Lagercrantz, U., Ellegren, H., and Andersson, L. (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Res.* **21:** 1111-1115.

116. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck, K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352:** 77-79.

117. Lauterbach, F., Kortner, C., Albracht, S.P., Unden, G., and Kroger, A. (1990) The fumarate reductase operon of Wolinella succinogenes. Sequence and expression of the frdA and frdB genes. *Arch. Microbiol.* **154:** 386-393.

118. Lawrence, J.G., and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44:** 383-397.

119. Levinson, G., and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4:** 203-221.

120. Li, L., Bannantine, J.P., Zhang, Q., Amonsin, A., May, B.J., Alt, D., Banerji, N., Kanjilal, S., and Kapur V. (2004) The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 12344-12349.

121. Liquori, C.L., Ricker, K., Moseley, M.L., Jacobsen, J.F., Kress, W., Naylor, S.L., Day, J.W., and Ranum, L.P. (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* **293:** 864-867.

122. Litt, M., and Luty, J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44:** 397-401.

123. Long, M. (2000) A new function evolved from gene fusion. *Genome Res.* **10:** 1655-1657.

124. Ma, C., and Redfield, R.J. (2000) Point mutations in a peptidoglycan biosynthesis gene cause competence induction in Haemophilus influenzae. *J. Bacteriol.* **182:** 3323-3330.

125. Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., et al. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33:** D192-D196.

126. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

127. Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W., and Vogelstein, B. (1995) Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268:** 1336-1338.

128. Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., and Moxon, E.R. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 3800-3804.

129. Matsuura, T., Yamagata, T., and Burgess, D.L. (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.* **26:** 191-194.

130. McAdam, R.A., Quan, S., Smith, D.A., Bardarov, S., Betts, J.C., Cook, F.C., Hooker, E.U., Lewis, A.P., Woollard, P., Everett, M.J., et al. (2002) Characterization of a Mycobacterium tuberculosis H37Rv transposon library reveals insertions in 351 ORFs and mutants with altered virulence. *Microbiology* **148:** 2975-2986.

131. McGuffin, L.J., Bryson, K., and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404-405.

132. McKinney, J.D., Honer zu Bentrup, K., Munoz-Elias, E.J., Miczak, A., Chen, B., Chan, W.T., Swenson, D., Sacchettini, J.C., Jacobs Jr, W.R., and Russell, D.G. (2000) Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* **406:** 683-685.

133. Meloni, R., Albanese, V., Ravassard, P., Treilhou, F., and Mallet, J. (1998) A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Hum. Mol. Genet.* **7:** 423-428.

134. Messier, W., Li, S.H., and Stewart, C.B. (1996) The birth of microsatellites. *Nature* **381:** 483.

135. Mizrahi, V., and Andersen, S.J. (1998) DNA repair in Mycobacterium tuberculosis. What have we learnt from the genome sequence? *Mol Microbiol* **29:** 1331-1339.

136. Morgante, M., and Olivieri, A.M. (1993) PCR-amplified microsatellites as markers in plant genetics. *Plant J.* **3:** 175-182.

137. Morgante, M., Hanafey, M., and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* **30:** 194-200.

138. Moxon, E.R., Rainey, P.B., Nowak, M.A., and Lenski, R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4:** 24-33.

139. Murphy, G.L., Connell, T.D., Barritt, D.S., Koomey, M., and Cannon, J.G. (1989) Phase variation of gonococcal protein II: regulation of gene expression by slipped strand mispairing of a repetitive DNA sequence. *Cell* **56:** 539-547.

140. Murray, V., Monchawin, C., and England, P.R. (1993) The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. *Nucleic Acids Res.* **21:** 23952398.

141. Nagafuchi, S., Yanagisawa, H., Ohsaki, E., Shirayama, T., Tadokoro, K., and Inoue, T. (1994) Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). *Nat. Genet.* **8:** 177-182.

142. Nauta, M.J., and Weissing, F.J. (1996) Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* **143:** 1021-1032.

143. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299-304.

144. OHearn, E., Holmes, S.E., Calvert, P.C., Ross, C.A., and L., M.R. (2001) SCA-12: Tremor with cerebellar and cortical atrophy is associated with a CAG repeat expansion. *Neurology* **56:** 299-303.

145. Ohta, T., and Kimura, M.A. (1973) Model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* **22:** 201204.

146. Patthy, L. (1996) Exon shuffling and other ways of module exchange. *Matrix Biol.* **15:** 301-310.

147. Peak, I.R.A., Jennings, M.P., Hood, D.W., Bisercic, M., and Moxon, E.R. (1996) Tetrameric repeat units associated with virulence factor phase variation in Haemophilus also occur in Neisseria spp. and Moraxella catarrhalis. *FEMS Microbiol. Lett.* **137:** 109-114.

148. Pearson, C.E., and Sinden, R.R. (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Current Op.Struct. Biol* **8:** 321-330.

149. Petes, T.D., Greenwell, P.W., and Dominska, M. (1997) Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* **146:** 491-498.

150. Pevzner, P.A., Borodovsky, M., and Mironov, A. A. (1989a) Linguistics of Nucleotide Sequences: II. Stationary Words in Genetic Texts and Domain Structure of DNA Sequence. *J. Biomol. Struc. Dyn* **6:** 1027-1038.

151. Pevzner, P.A., Borodovsky, M., and Mironov, A.A. (1989b) Linguistics of Nucleotide Sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dyn.* **6:** 1013-1026.

152. Pevzner, P.A., Borodovsky, M., and Mironov, A.A. (1989) Linguistics of Nucleotide Sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dyn.* **6:** 1013-1026.

153. Ponting, C.P., and Russell, R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31:** 45-71.

154. Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1992) Numerical Recipes in C. Cambridge: Cambridge University Press.

155. Primmer, C.R., Mller A.P., Ellegren, H. (1996) A wide-ranging survey of cross-species amplification in birds. *Mol. Ecol.* **5:** 365 378.

156. Primmer, C.R., and Ellegren, H. (1998) Patterns of molecular evolution in avian microsatellites. *Mol. Biol. Evol.* **15:** 9971008.

157. Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Mller, A.P., and Ellegren, H. (1997) Low frequency of microsatellites in the avian genome. *Genome Res.* **7:** 471-482.

158. Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J.C., and Perucho, M. (1997) Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science* **275:** 967-969.

159. Ranum, L.P.W., Moseley, M.L., Leppert, M.F., van den Engh, G., La Spada, A.R., and Koob, M.D. (1999) Massive CTG expansions and deletions may reduce penetrance of spinocerebellar ataxia type 8. *Am. J. Hum. Genet.* **65:** A466.

160. Raynaud, C., Laneelle, M., Ryan, H.S., Draper, P., Laneelle, G., and Daffe, M. (1999) Mechanisms of pyrazinamide resistance in mycobacteria: importance of lack of uptake in addition to lack of pyrazinamidase activity. *Microbiology* **145:** 1359-1367.

161. Richards, R.I. (2001) Dynamic mutations: a decade of unstable expanded repeats in human genetic disease. *Hum. Mol. Genet.* **10:** 2187-2194.

162. Richards, R.I., and Sutherland, G.R. (1992) Dynamic mutations: a new class of mutations causing human disease. *Cell* **70:** 709-712.

163. Ritz, D., Lim, J., Reynolds, C.M., Poole, L.B., and Beckwith, J. (2001) Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion. *Science* **294:** 158-160.

164. Robertson, B.D., and Mayer, T.F. (1992) Genetic variation in pathogenic bacteria. *Trends Genet.* **8:** 422-427.

165. Rocha, E.P.C., Pradillon, O., Bui, H., Sayada, C., and Denamur, E. (2002) A new family of highly variable proteins in the Chlamydophila pneumoniae genome. *Nucleic Acids Res.* **30:** 4351-4360.

166. Roche, R.J., and Moxon, E.R. (1995) Phenotypic variation in H. influenzae:the interrelationship of colony opacity, capsule and lipopolysaccharide. *Microb. Pathog.* **18:** 129-140.

167. Rosenberg, S.M., Longerich, S., Gee, P., and Harris, R.S. (1994) Adaptive mutation by deletions in small mononucleotide repeats. *Science* **265:** 405-407.

168. Rose, O., and Falush, D. (1998) A threshold size for microsatellite expansion. *Mol. Biol. Evol.* **15:** 613-615.

169. Rousseau, C., Sirakova, T.D., Dubey, V.S., Bordat, Y., Kolattukudy, P.E., Gicquel, B., and Jackson, M. (2003) Virulence attenuation of two Mas-like polyketide synthase mutants of Mycobacterium tuberculosis. *Microbiology* **149:** 1837-1847.

170. Santibez-Koref, M.F., Gangeswaran, R., and Hancock, J.M. (2001) A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol. Biol. Evol.* **18:** 2119-2123.

171. Sarkari, J., Pandit, N., Moxon, E.R., and Achtman, M. (1994) Variable expression of the Opc outer membrane protein in Neisseria meningitidis is caused by size variation of a promoter containing poly-cytidine. *Mol. Microbiol.* **13:** 207-217.

172. Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48:** 77-84.

173. Saunders, N.J., Jeffries, A.C., Peden, J.F., Hood, D.H., Tettelin, H., Rappuoli, R., and Moxon, E.R. (2000) Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. *Mol. Microbiol.* **37:** 207-215.

174. Schlotterer, C. (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109:** 365-371.

175. Schlotterer, C., and Tautz, D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20:** 211-215.

176. Schug, M.D., Wetterstrand, K.A., Gaudette, M.S., Lim, R.H., Hutter, C.M., and Aquadro, C.F. (1998) The distribution and frequency of microsatellite loci in Drosophila melanogaster. *Mol. Ecol.* **7:** 57-70.

177. Seraphin, B. (1992) The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals. *DNA Seq.* **3:** 177-179.

178. Sharp, P.M., Shields, D.C., Wolfe, K.H., and Li, W.H. (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246:** 808-810.

179. Shinde, D., Lai, Y., Sun, F., and Arnheim, N. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.* **31:** 974980.

180. Shriver, M.D., Jin, L., Chakraborty, R., and Boerwinkle, E. (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* **134:** 983993.

181. Springer, B., Sander, P., Sedlacek, L., Hardt, W., Mizrahi, V., Schr, P., and Bttger, E.C. (2004) Lack of mismatch correction facilitates genome evolution in mycobacteria. *Mol. Microbiol.* **53:** 1601-1609.

182. Sreenu, V.B., Alevoor, V., Nagaraju, J., and Nagarajaram, H.A. (2003a) MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res* **31:** 106-108.

183. Sreenu, V.B., Ranjitkumar, G., Swaminathan, S., Priya, S., Bose, B., Pavan, M.N., Thanu, G., Nagaraju, J., and Nagarajaram, H.A. (2003b) MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. *Appl. Bioinformatics* **2:** 165-168.

184. Stephan, W., and Cho, S. (1994) Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics* **136:** 333-341.

185. Stern, A., and Meyer, T.F. (1987) Common mechanism controlling phase and antigenic variation in pathogenic neisseriae. *Mol. Microbiol.* **1:** 5-12.

186. Stern, A., Brown, M., Nickel, P., and F., M.T. (1986) Opacity genes in N. gonorrhoeae: control of phase and antigenic variation. *Cell* **47:** 61-71.

187. Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365:** 274276.

188. Strassmann, J.E., Barefield, K., Solis, C.R., Hughes, C.R., and Queller, D.C. (1997) Trinucleotide microsatellite loci for a social wasp, Polistes. *Mol. Ecol.* **6:** 97-100.

189. Streelman, J.T., and Kocher, T.D. (2002) Microsatellite variation associated with prolactin expression and growth of saltchallenged Tilapia. *Physiol. Genomics* **9:** 1-4.

190. Su, T.Z., Schweizer, H.P., and Oxender, D.L. (1991) Carbon-starvation induction of the ugp operon, encoding the binding protein-dependent sn-glycerol-3-phosphate transport system in Escherichia coli. *Mol. Gen. Genet.* **230:** 28-32.

191. Tanaka, N., Kinoshita, T., and Masukawa, H. (1968) Mechanism of protein synthesis inhibition by fusidic acid and related antibiotics. *Biochem. Biophys. Res. Commun.* **30:** 278-283.

192. Tautz, D. (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. Basel: Birkhauser.

193. Theiss, P., and Wise, K.S. (1997) Localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a mycoplasma ABC transporter operon. *J. Bacteriol.* **179:** 4013-4022.

194. Thibodeau, S.N., Bren, G., and Schaid, D. (1993) Microsatellite instability in cancer of the proximal colon. *Science* **260:** 816-819.

195. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680.

196. Tidow, N., Boecker, A., Schmidt, H., Agelopoulos, K., Boecker, W., Buerger, H., and Brandt, B. (2003) Distinct amplification of an untranslated regulatory sequence in the egfr gene contributes to early steps in breast cancer development. *Cancer Res.* **63:** 1172-1178.

197. Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **388:** 539-547.

198. Toth, G., Gaspari, Z., and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10:** 967981.

199. Usdin, K., and Grabczyk, E. (2000) DNA repeat expansions and human disease. *Cell. Mol. Life Sci.* **57:** 914-931.

200. Valdes, A.M., Slatkin, M., and Freimer, N.B. (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133:** 737749.

201. Valvano, M.A., Messner, P., and Kosma, P. (2002) Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. *Microbiology* **148:** 1979-1989.

202. van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62:** 275-293.

203. Van Ham, S.M., van Alphen, L., Mooi, F.R., and van Putten, J.P.M. (1993) Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* **73:** 1187-1196.

204. Van Ham, S.M., van Alphen, L., Mooi, F.R., and van Putten, J.P.M. (1994) The fimbrial gene cluster of H. influenzae type b. *Mol. Microbiol.* **13:** 673-684.

205. van Soolingen, D., de Haas, P.E.W., Hermans, P.W.M., Groenen, P.M.A., and van Embden, J.D.A. (1993) Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. *J. Clin. Microbiol* **31:** 1987-1995.

206. Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., and Pizzuti, A. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65:** 905-914.

207. Waddell, S.J., Chung, G.A., Gibson, K.J.C., Everett, M.J., Minnikin, D.E., Besra, G.S., and Butcher, P.D. (2005) Inactivation of polyketide synthase and related genes results in the loss of complex lipids in Mycobacterium tuberculosis H37Rv. *Lett. Appl. Microbiol.* **40:** 201-206.

208. Wayne, L.G., and Hayes, L. (1996) An in vitro model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence. *Infect. Immun.* **64:** 2062-2069.

209. Weber, J.L. (1990) Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7:** 524-530.

210. Weber, J.L., and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388-396.

211. Weiser, J.N., Love, J.M., and Moxon, E.R. (1989) The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. *Cell* **59:** 657-665.

212. Weiser, J.N., Maskell, D.J., Butler, P.D., Lindberg, A.A., and Moxon, E.R. (1990) Characterization of repetitive sequences controlling phase variation of Haemophilus inflenzae lipopolysaccharide. *J. Bacteriol.* **172:** 3304-3309.

213. Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A, and Tingey, S.V., (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18:** 6531-6535.

214. Wolfa, D.M., Vazirani, V.V., and Arkin, A.P. A microbial modified prisoners dilemma game: how frequencydependent selection can lead to random phase variation. J. Theor. Biol: in press.

215. Wren, J.D., Forgacs, E., Fondon 3rd, J.W., Pertsemlidis, A., Cheng, S.Y., Gallardo, T., Williams, R.S., Shohet, R.V., Minna, J.D., and Garner, H.R. (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.* **67:** 345-356.

216. Yamada, T., Koyama, T., Ohwada, S., Tago, K., Sakamoto, I., Yoshimura, S., Hamada, K., Takeyoshi, I., and Morishita, Y. (2002) Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. *Cancer Lett.* **181:** 115-120.

217. Yang, J., Wang, J., Chen, L., Yu, J., Dong, J., Yao, Z.J., Shen, Y., Jin, Q., and Chen, R. (2003) Identification and characterization of simple sequence repeats in the genomes of Shigella species. *Gene* **322:** 85-92.

218. Yogev, D., Rosengarten, R., Watson, R., and Wise, K.S. (1991) Molecular basis of Mycoplasma surface antigenic variation: a novel set of divergent

genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences. *EMBO J.* **10:** 4069-4079.

219. Youings, S.A., Murray, A., Dennis, N., Ennis, S., Lewis, C., McKechnie, N., Pound, M., Sharrock, A., and Jacobs, P. (2000) FRAXA and FRAXE: the results of a five year survey. *J. Med. Genet.* **37:** 415-421.

220. Yu, M.W., Yang, Y.C., and Yang, S.Y. (2002) Androgen receptor exon 1 CAG repeat length and risk of hepatocellular carcinoma in women. *Hepatology* **36:** 156-163.

221. Zhu, Y., Queller, D.C., and Strassmann, J.E. (2000) A phylogenetic perspective on sequence evolution in microsatellite loci. *J. Mol. Evol.* **50:** 324-338.

222. Zhu, Y., Strassmann, J.E., and Queller, D.C. (2000) Insertions, substitutions, and the origin of microsatellites. *Genome Res.* **76:** 227236.

223. Zoghbi, H.Y., and Orr, H.T. (2000) Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.* **23:** 217-237.