Analysis of simple sequence repeats in genome and protein sequences and development of computational tools for comparative promoter sequence analysis

A thesis submitted to the University of Pune
for the Degree of

Doctor of Philosophy

in

Biotechnology

By

Mukund Vyankatesh Katti

Plant Molecular Biology Unit
Division of Biochemical Sciences
National Chemical Laboratory
Pune 411008, INDIA

April 2001

CONTENTS

ACKNOWLEDGEMENTS

It gives me immense pleasure to express my gratitude towards my research guide Dr. Vidya Gupta. I am grateful to her for all the advice, guidance, support and encouragement during every stage of this work. I thank her for giving me the freedom to explore into this particular topic of my interest.

I express my sincere thanks to Dr. P. K. Ranjekar for his warm-hearted support and motivation. Several discussions with him have given me valuable ideas and have helped in planning the work.

I would like to thank Prof. M. V. Hegde, Dr. C. G. Suresh and Dr. V. Shankar for suggestions on the manuscripts. I am grateful to Prof. M. R. N. Murthy for encouragement and for giving me an opportunity to visit his laboratory at IISc, Bangalore. I thank Director, Bioinformatics Centre, University of Pune, Pune for providing me the library facilities.

My special thanks are due to Dr. Premnath for his support in making available the *TRIPS* and *SSR* databases through NCL web site. Help from Dr. Meena Sakharkar has been instrumental during implementation of *TRES* program at the National University of Singapore web server. I gratefully acknowledge her help and interest in *TRES*. I also wish to thank Dr. Edger Wingender, Dr. David Ghosh and Dr. Kenichi Higo for allowing use of site libraries from *TRANSFAC*, *ooTFD* and *PLACE* databases, respectively, in *TRES* program.

I do not have enough words to express my gratitude towards my friends Sami, Bhushan, Vrinda, Rajesh, Rao and Dr. Ashok Aspatwar for their cheerful company and for being with me at all the times.

I thank Mr. B. G. Patil, Dr. Mohini Sainani, Dr. Meena Lagu, Dr. Nirmala, Dr. Lalitha, Dr. Shubhada and Indira for their help from time to time. I also thank Mr. Karunakaran and Mr. Jagtap for their assistance.

I am thankful to all my colleagues Ajit, Abhay, Ashok, Aditi, Ajay, Anjali, Aparna, Archana, Armaity, Arundhati, Bimba, Deepak, Jakir, Manoj, Milind, Meena, Maneesha, Manisha, Raju, Rahul, Rajashekhar, Renu, Renuka, RK, Sadhana, Sanjay, Sastry, Shashi, Suresh, Suvarna, Swati, Venkat and Vijay for their cooperation and friendly help in the laboratory.

I gratefully acknowledge Council of Scientific and Industrial Research, New Delhi for the award of CSIR Junior and Senior Research Fellowship. I would also like to thank Director, NCL for providing all the necessary facilities during my tenure at NCL.

In the end, I remain indebted to my parents and all my family members for their care, love, support and encouragement.

Mukund V. Katti

DECLARATION

Certified that the work incorporated in the thesis, entitled "Analysis of simple sequence repeats in genome and protein sequences and development of computational tools for comparative promoter sequence analysis", submitted by Mr. Mukund V. Katti, was carried out by the candidate under my supervision. Such material as has been obtained from other sources has been duly acknowledged in the thesis.

Vidya S. Gupta
(Research Guide)

THESIS ABSTRACT

During the past few years, there has been exponential growth in the availability of biomolecular sequence information. Though, the DNA and protein sequences contain important biological information it can be rationalized only by careful computational analysis. Some of the important areas where computer analyses have played a significant role in new biological discovery include functional annotation of DNA and protein sequences, sequence alignment, phylogenetic analysis, identification of novel genes in genome sequences, and prediction of protein structure. With some practical knowledge of computer programming, I realized that availability of large amount of sequence data would provide opportunities to generate new information. In my thesis, I have applied computer programming to analyze the biomolecular sequences to address some specific questions that occurred to me.

Important findings of my work:

[1] Differential distribution of simple sequence repeats in eukaryotic genome sequences:

Simple sequence repeats are ubiquitous in eukaryotic genome sequences and are thought to contribute in genome organization and evolution. Availability of complete genome sequences now allows determination of the extent to which repeats are generated in a genome. Therefore, I analyzed complete chromosome sequences available from human, *Drosophila*, *C. elegans*, *Arabidopsis* and yeast to assess the occurrence of mono-, di-, tri-, and tetranucleotide repeats at whole genome/chromosome level.

In all the genomes studied, dinucleotide repeats seem to be longer compared to other repeats. Additionally, tetranucleotide repeats in human and tri-nucleotide repeats in *Drosophila* also tend to be long. Although, the trends for different repeat classes are similar between different chromosomes within a genome, the density of repeats may vary between different chromosomes of a species. Abundance or rarity of various di- and trinucleotide repeat classes in different genomes could not be explained by nucleotide composition of a sequence or potential of repeated motifs to form alternative DNA structures. This suggests that in addition to nucleotide composition of repeat motifs, characteristic DNA replication/repair/recombination machinery might play an important role in the genesis of repeats.

I also examined the occurrences of codon (trinucleotide) repeats in all the predicted coding DNA sequences of *Drosophila*, *C. elegans* and yeast genomes. My study has revealed that codon repeats corresponding to small hydrophilic amino acids are more frequent compared to codon repeats encoding hydrophobic amino acids.

Based on my analysis, I have developed a web-resource on simple sequence repeats in eukaryotic genome sequences and complete genome coding DNA sequences, which is available at the URL: http://www.ncl-india.org/ssr. This resource could be useful to identify a wide range of microsatellite loci to study their sequence and position dependent evolution.

[2] Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications:

Though simple sequence repeats originate due to errors during DNA processing, when they occur in the coding regions it might lead to appearance of repeated sequence patterns in protein sequences. I have analyzed a large collection of protein sequences from the SWISS-PROT database to assess how common are the internal repeats in proteins and what are their implications on protein structure and function.

My study shows that single amino acid repeats of small hydrophilic amino acids like glutamine, serine, glutamic acid, glycine, and alanine are more frequent in proteins compared to repeats of hydrophobic amino acids. However, the regions containing tandem single amino acid repeats do not seem to be clearly assigned to any functional domains. A few examples of single amino acid repeats in solved structures indicate that these regions may adopt regular as well as non-regular structures and this could be largely influenced by their context in parent proteins.

Tandem oligo-peptide repeats of different types with varying levels of conservation have been detected in several proteins and found to be conspicuous particularly in structural and cell surface proteins. Available structural studies suggest that repeated sequence patterns can lead to repeated structural patterns in proteins, particularly when repeating units are longer (>20 residues). However, we still do not know much about the structures formed by short tandem repeats. It appears that repeated sequence patterns may be a mechanism that provides regular arrays of spatial and functional

groups, useful for structural packing or for one to one interactions with target molecules.

I have compiled the results of the above analysis in the form of a database of *Tandem Repeats in Protein Sequences (TRIPS)* which is available at the URL: http://www.ncl-india.org/trips. The *TRIPS* database gives a systematic and comprehensive picture of repeat patterns observed in protein sequences and could be useful for further explorations.

[3] Development of a web based software tool, TRES, for comparative promoter sequence analysis

Computational search of promoter DNA sequences helps to identify putative sequence motifs possibly involved in transcription regulation. Rather than searching a single sequence, simultaneous analysis of several related sequences can be more informative and useful to identify common regulatory modules conserved in a set of sequences. Considering the effectiveness of this approach, I have developed a web based software tool TRES (Transcription Regulatory Element Search) that allows simultaneous analysis of as many as 20 promoter sequences for putative regulatory elements. TRES has been organized in 4 analysis tools, namely (1) Matrix-search (2) IUPAC-string search, (3) Palindrome search and (4) k-tuple search. TRES is implemented on a web-server at the URL: http://bioportal.bic.nus.edu.sg/tres. This interactive web interface enables the user to select program module, choose search parameters and submit the sequences for online search.

The advantage of TRES over other available programs is that several related sequences can be analyzed simultaneously and putative motifs conserved in all or in majority of the sequences can be identified. Thus, motifs that occur only in one or a few sequences, possibly due to chance, can be filtered. TRES could be used to identify evolutionarily conserved motifs in orthologous sequences. It can be also used to elucidate common regulatory modules in genes that show similar patterns of expression. With ever-increasing sequence information available from diverse species, comparative promoter analysis appears to be a promising strategy to identify regulatory modules in genes of interest.

My thesis has been organized in five chapters and highlights of the contents in each chapter are as follows:

Chapter 1: Introduction:

I have briefly reviewed how sequence information has grown exponentially and how computational analyses of nucleic acid and protein sequences helps in understanding biological principles. The genesis of thesis and objectives of the thesis are also included here.

Chapter 2: Differential distribution of simple sequence repeats in eukaryotic genome sequences:

In this chapter, I have presented my findings from the analysis of simple sequence repeats in complete genome/chromosome sequences available from a few eukaryotic genomes. In addition, occurrence of codon repeats in complete genome coding DNA sequences of *Drosophila*, *C. elegans* and yeast are also analyzed.

Chapter 3: Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications:

Here, I have included results of my studies on occurrences of internal repeats in protein sequences based on the analysis of SWISS-PROT protein sequence database.

Chapter 4: Development of a web based software tool, TRES, for comparative promoter sequence analysis:

In this chapter, I have described development of the TRES program, its advantages and possible applications.

Chapter 5: Thesis Overview:

Here, I have briefly summarized important findings of my work and future perspectives.

The list of references is given at the end.

LIST OF ABBREVIATIONS

aa	amino acid
bp	base pairs
cDNA	complimentary DNA
DNA	Deoxyribose Nucleic Acid
EST	Expressed Sequence Tag
ftp	file transfer protocol
HTML	Hyper-Text Markup Language
http	hyper-text transfer protocol
IUPAC	International Union of Pure and Applied Chemistry
kbp	kilo base pairs
ln	natural logarithm
mbp	million base pairs
mRNA	messenger RNA
nt	nucleotides
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PDB-ID	Protein Data Bank Identifier
RNA	Ribose Nucleic Acid
SSR	Simple Sequence Repeats
TF	Transcription Factor
TRES	Transcription Regulatory Element Search (program)
TRIPS	Tandem Repeats In Protein Sequences (database)
URL	Uniform Resource Locator
WWW	World Wide Web

IUPAC single letter codes for nucleotide bases:

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Adenine or Guanine
Y	Cytosine or Thymine
S	Cytosine or Guanine
W	Adenine or Thymine
K	Guanine or Thymine
M	Adenine or Cytosine
B	Cytosine or Guanine or Thymine
D	Adenine or Guanine or Thymine
H	Adenine or Cytosine or Thymine
V	Adenine or Cytosine or Guanine
N or X	any or unknown base

Single letter codes for amino acid residues:

A	Alanine	C	Cysteine	D	Aspartic acid
E	Glutamic acid	F	Phenylalanine	G	Glycine
H	Histidine	I	Isoleucine	K	Lysine
L	Leucine	M	Methionine	N	Asparagine
P	Proline	Q	Glutamine	R	Arginine
S	Serine	T	Threonine	V	Valine
W	Tryptophan	Y	Tyrosine	X	any or unknown residue

INTRODUCTION:

Computer applications in nucleic acid and protein sequence analysis

---

Every living organism is empowered with all the information necessary for its growth, development, maintenance, and reproduction. Among various macromolecules present in the cells, DNA acts as a carrier of genetic information whereas RNA and proteins allow expression of this information into function. The cellular factors read the information in the DNA and build various components that catalyze all the chemical reactions occurring in a cell, sense changes in the environment, interact with each other and self assemble to form complex cellular machineries. This *Central Dogma of Life* has emerged as an unifying theme in all biological systems.

Since the nucleotide sequence of a DNA or amino acid sequence of a protein determines its function, knowledge of their sequence is imperative for understanding life processes at molecular level. Naturally, DNA and protein sequencing has received great importance in biological research which is evident from the amount of sequence data accumulated during past few years. In this chapter, I have attempted to briefly review how sequence information has grown exponentially and how its analysis helps in discovery of new biological principles.

1.1 DNA sequencing: From genes to genomes

Determination of nucleotide sequence of a DNA involves establishing chemical identities of successive nucleotides in a specific DNA region of interest. Efficient DNA sequencing methods were developed as early as 1977 employing base specific chemical cleavage (Maxam and Gilbert, 1977) or chain termination using specific di-deoxy nucleotides in enzymatic DNA synthesis (Sanger et al., 1977b). Subsequently, automated DNA sequencers were designed that greatly reduced cost and labor in DNA sequencing (Smith et al., 1986).

For efficient DNA sequencing, it is essential to obtain a large amount of homogenous preparation of DNA which is possible through cloning of desired region of DNA into a suitable vector and its subsequent transformation into host cells that allows rapid multiplication and purification. The first step in cloning of a gene involves preparation of

genomic or cDNA libraries which are then screened using oligo-nucleotides designed from partial amino-acid sequence of a purified protein or using a homologous gene-probe from related species. Alternatively, cDNA expression libraries can be screened using antibodies developed against a purified protein of interest. Polymerase Chain Reaction (PCR) provides another rapid method to amplify DNA region of interest that can be cloned or sequenced directly.

With the availability of automated DNA sequencers, Adams et al., (1991) suggested that a large number of randomly selected cDNA clones could be sequenced economically and this could be a fast approach to discover new genes. They termed these sequences as Expressed Sequence Tags (EST) and showed that ESTs could also be used as markers to construct saturated genome maps and identify coding regions in genomic sequences. This approach has become popular rapidly and thousands of ESTs from different tissues of a wide range of organisms have been obtained during the past one decade.

Since the DNA carries all the genetic information, it was realized that deciphering complete genome sequence of a species could be a major step in understanding biological processes at molecular level. For complete genome sequencing of a microbial or an eukaryotic genome, two different strategies have been used: clone by clone approach and whole genome random shotgun sequencing. In clone by clone approach, first large insert libraries representing whole genome or a single chromosome are constructed followed by mapping of individual clones to obtain overlapping series of clones spanning an entire chromosome or a genome. These mapped clones are then sequenced one by one to obtain a complete sequence. Venter and his colleagues (Fleischmann et al., 1995) suggested whole genome random shotgun sequencing and computer assisted assembly of sequences to obtain a complete genome sequence. Statistical analysis indicates that when whole genomes are randomly sequenced in sufficient excess (6X to 8X fold) more than 99.9% sequence coverage can be obtained. However, sophisticated computer tools are required to align and overlap a large number of single pass random sequences. A combination of small-insert (~2 kbp) and large-insert (~15-20 kbp) libraries helps to obtain overlaps across large segments. After final assembly, sequence gaps are closed following clone by

clone approach. The advantages of the shotgun method are that initial efforts in mapping large-insert clones are not required and with the help of automated sequencers the cost per sequencing of a base can be markedly reduced.

The first genome ever sequenced completely was that of bacteriophage ϕX-174 (Sanger et al., 1977a). Later, complete sequence of the first mitochondrial genome from human (Anderson et al., 1981) and the first chloroplast genome from tobacco (Shinozaki et al., 1986) were obtained. In 1995, Fleischmann et al., were able to sequence complete genome of a bacterium *Haemophilus influenzae* applying whole genome random shotgun sequencing. That was the first free living organism to be sequenced completely, which marked the beginning of a new era of genomics (Table 1.1). Subsequently, complete genome sequencing of the first eukaryote, yeast (Goffeau et al., 1996) and the first multicellular organism, *C. elegans* (The *C. elegans* Sequencing Consortium, 1998) were achieved. The greatest accomplishment of these efforts has been unraveling of the complete DNA sequence of the human genome (Venter et al. 2001; International Human Genome Sequencing Consortium, 2001). Indeed, it is an astonishing achievement for biologists that within a span of a decade complete genomes of human, 4 other eukaryotes, 9 archea, 36 bacteria, several viruses and organelles have been sequenced, with many more genomes in pipeline.

1.2 The expanding universe of sequence databases:

For rapid dissemination of sequence information and their maximal utilization, researchers deposit their newly sequenced data in the central databases like GenBank, DNA Databank of Japan (DDBJ), and/or EMBL Nucleotide Sequence Database that accept, curate, maintain and distribute DNA and/or protein sequence information. Due to advances in recombinant DNA technology and automated DNA sequencers, sequencing has become an easy and highly productive effort for researchers. Consequently, the amount of available sequence information has increased exponentially during the past two decades (Table 1.2). Moreover, high throughput sequencing efforts to obtain expressed sequence tags and complete genome sequences continue to flood the sequence

databases. Concurrently, the spread of Internet throughout the world has revolutionized the way data can be disseminated and retrieved. Now, any researcher connected to the worldwide electronic network through a computer can access up-to-date sequence information anytime and anywhere in the world. The availability of search engines and sequence analysis tools on the web have allowed rapid analysis of data to obtain meaningful information and help researchers in designing experiments.

Table 1.1 Important milestones in genome sequencing

Organism	Genome size (bp)	Year of completion	Reference
Phage / Organelles			
Bacteriophage: ϕX-174	5,386	1977	Sanger et al., (1977a)
Bacteriophage lambda	48,502	1982	Sanger et al., (1982)
Human mitochondrion	16,569	1981	Anderson et al., (1981)
Tobacco chloroplast	155,939	1986	Shinozaki et al., (1986)
Bacteria / Archea			
Haemophilus influenzae	1,830,137	1995	Fielschmann et al., (1995)
Mycoplasma genitalium	580,070	1995	Fraser et al., (1995)
Methanococcus jannaschii	1,664,970	1996	Bult et al., (1996)
Bacillus subtilis	4,214,814	1997	Kunst et al., (1997)
Escherichia coli	4,639,221	1997	Blattner et al., (1997)
Mycobacterium tuberculosis	4,411,529	1998	Cole et al., (1998)
Vibrio cholerae	4,033,460	2000	Heidelberg et al., (2000)
Eukaryotes			
Saccharomyces cerevisiae	~12,068,000	1996	Goffeau et al., (1996)
Caenorhabditis elegans	~95,530,000	1998	The *C. elegans* sequencing consortium (1998)
Drosophila melanogaster	~120,000,000	2000	Adams et al., (2000)
Arabidopsis thaliana	115,409,949	2000	The Arabidopsis Genome Initiative (2000)
Human Genome			
Chromosome-22	33,618,270	1999	Dunham et al., (1999)
Chromosome-21	33,824,148	2000	Hattori et al., (2000)
Human genome draft sequence	~2,910,000,000	2001	Venter et al., (2001); International Human Genome Sequencing Consortium (2001)

Table 1.2: The growth of sequence data in GenBank*

Year	Number of Base Pairs	Number of Sequences
1982	680,338	606
1985	5,204,420	5,700
1990	49,179,285	39,533
1994	217,102,462	215,273
1996	651,972,984	1,021,211
1998	2,008,761,784	2,837,897
2000	11,101,066,288	10,106,023

*http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html (Jan. 2001)

1.3 Making sense from the sequence:

The sheer volume of sequence data makes use of computers inevitable in storage, curation, annotation, mining, dissemination and analysis of sequence information. Though DNA or protein sequences contain important biological information, it is hard to make any sense from merely examining a sequence. It is only by further careful computational analysis that any information contained in a sequence can be understood. Analysis of an individual DNA or protein sequence helps identify previously characterized sequence features that can give important insights into the structure and function of a gene or a protein. Besides, comparative analysis of a large collection of sequence data-set can be useful in discovering unifying principles in biology and understanding evolutionary relationships. However, before application of computational tools for information analysis, it is essential to know how information is stored in DNA, how genes are organized and expressed, what are the features of proteins, how they fold and how they function (Figures 1.1 and 1.2). Indeed, these principles should direct any computational analysis aimed at unraveling biological information hidden in DNA and protein sequences.

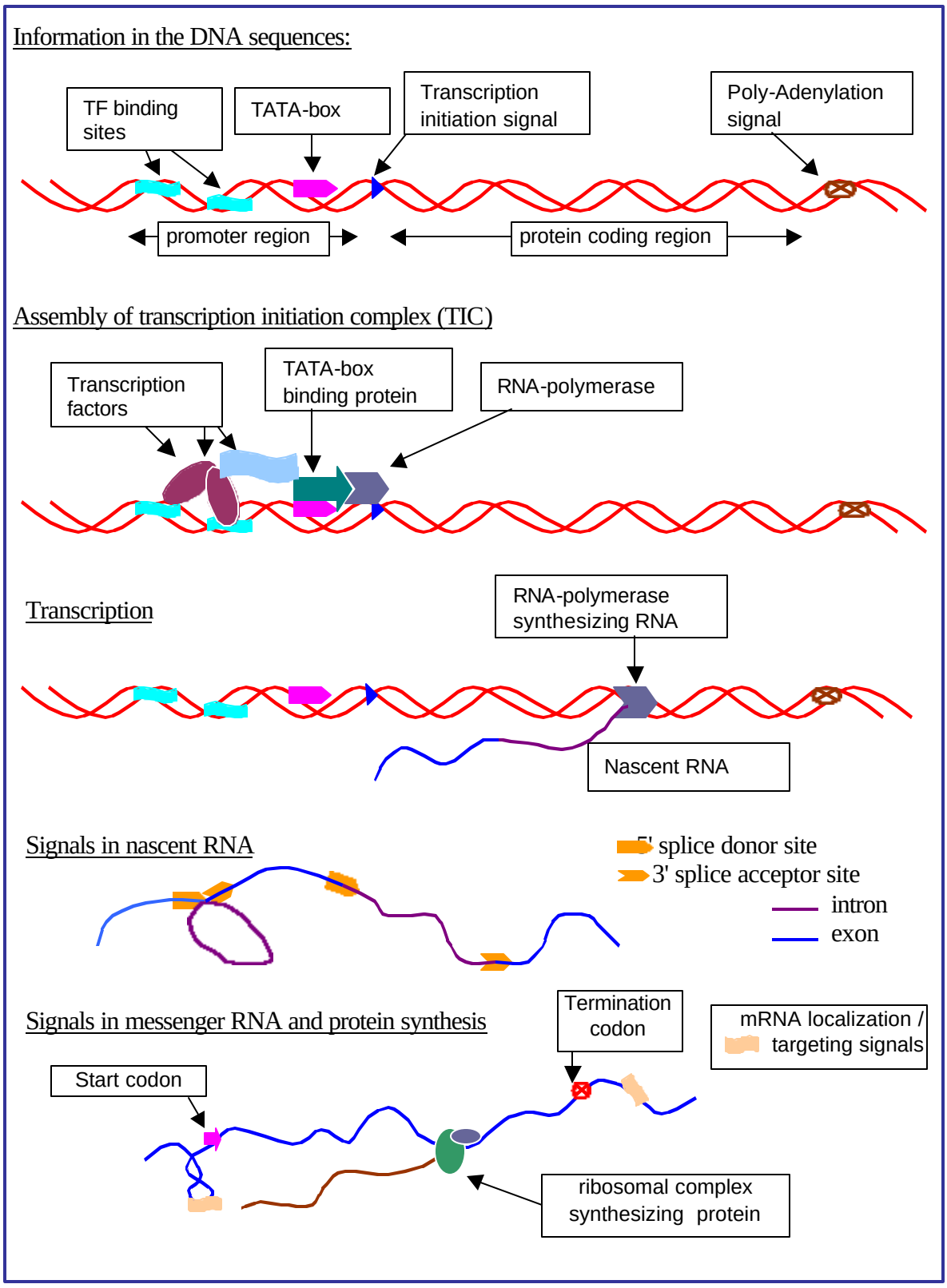


Figure 1.1: Information contained in the sequences of DNA and RNA molecules

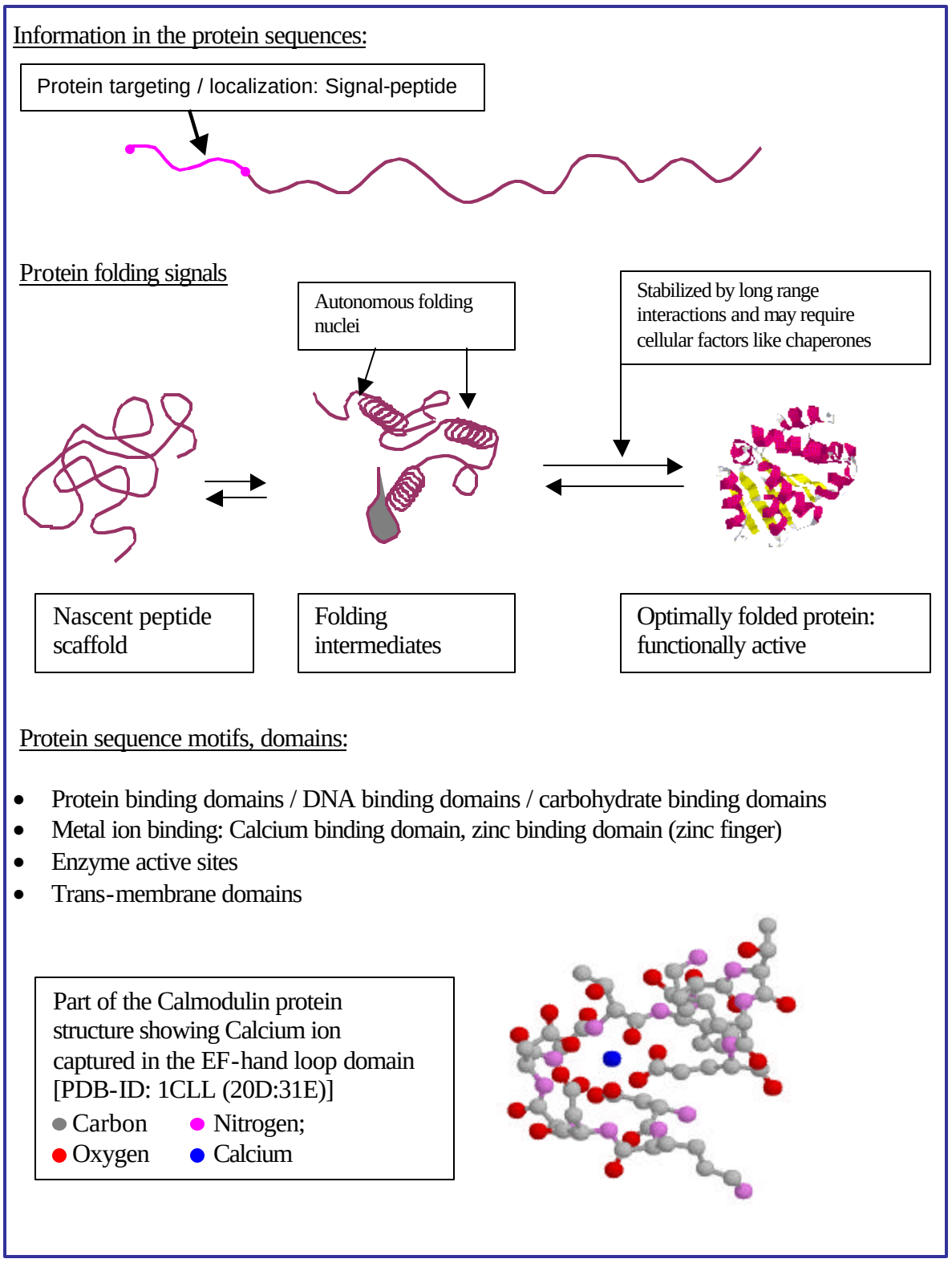


Figure 1.2: Information contained in the sequences of protein molecules

1.3.1 Finding homologous sequences by database search:

Whenever a new sequence is obtained, the first approach for its functional annotation is to check whether similar sequences have been studied earlier. If similar sequences are available, information about their function or structure can be extended to the new sequence. For example, when *Pto* resistance gene in tomato was mapped it was not known how it conferred resistance to the races of *Pseudomonas*. However, when this gene was later cloned, its deduced amino acid sequence revealed similarity to serine-threonine protein kinases, suggesting a role for *Pto* in signal transduction pathway triggering defense response (Martin et al., 1993). Thus homology search provides a powerful tool to assign function to new sequences obtained from genome sequencing, map based cloning, random cDNA sequencing (ESTs), differential display or subtractive hybridization.

Typically a database search program accepts a query sequence and aligns it with each and every sequence in the database and reports whenever it finds significant similarity. For pairwise sequence alignment, dynamic programming algorithms are guaranteed to find maximal similarity. However, they are highly time consuming and hence are not suitable for scanning large databases. Therefore, fast algorithms that seek local similarity are employed in database search programs. In contrast to global similarity algorithms that try to optimize overall alignment of two sequences, local similarity algorithms seek only relatively conserved subsequences that could be done very fast if gaps need not be considered. These methods are more useful particularly when comparing a cDNA sequence against a genomic sequence or when distantly related sequences share only isolated regions of similarity probably around functionally important regions.

Wilbur and Lipman (1983) proposed a rapid similarity search algorithm useful for database search. In the first step, locations of all the k-tuple matches between a query sequence and a database sequence are determined in a single pass. Further, rather than considering all k-tuple matches, only those occurring in a specified window on significant diagonals are compared and alignment calculated. With the optimal k-tuple

and window size, this program could rapidly search a large sequence database. Later, Lipman and Pearson (1985) and Pearson and Lipman (1988) implemented modified versions of the above algorithm as FASTA programs, which soon became popular due to their speed, selectivity and sensitivity. In the FASTA programs, initially a lookup table is used to locate all identities or group of identities (k-tuples) between two DNA or protein sequences. Next, the regions are scanned using a scoring matrix and the best regions above a certain threshold are saved. Then, the initial regions are joined optimally using a more rigorous dynamic programming algorithm to give most optimal alignment.

Altschul et al. (1990) used local alignment approach to device a highly rapid Basic Local Alignment Search Tool (BLAST). Briefly, BLAST tries to identify the highest scoring pairs of identical length segments between two sequences (MSP- Maximal Segment Pair). While scanning through a sequence, BLAST quickly determines whether the sequence contains a word of length 'w' that can pair with the query sequence with a score more than a certain threshold value 'T'. All such hits are extended to check if they can attain a minimum score of 'S'. Simulation studies by scanning of random sequences against a test database, have allowed determination of optimal values for 'w', 'T' and 'S' and have also permitted to assign statistical significance to high scoring MSPs detected by BLAST.

In further modification of BLAST, rather than considering single hits with higher threshold 'T', simultaneous occurrences of two non-overlapping hits with lesser threshold 'T' are considered (Altschul et al., 1997). Non-overlapping MSPs can be then aligned allowing gap and yet time can be kept to a minimum. In the Position Specific Iterated BLAST (PSI-BLAST), statistically significant alignments produced by BLAST are combined to create a profile matrix, which is then used in additional round of database search to detect weak similarities. Many DNA or protein sequences contain small structurally or functionally important regions, domains or sequence patterns and researchers are interested to know other related sequences with similar patterns. To facilitate such analysis, Zhang et al., (1998) developed a Pattern Hit Initiated BLAST (PHI-BLAST) program wherein users can submit a sequence along with a specific pattern

contained in it. The PHI-BLAST first scans the database for sequences containing the pattern and then compares such sequences with query sequence using BLAST algorithm. A central BLAST database search facility available at http://www.ncbi.nlm.nih.gov/blast provides a convenient interface to search the DNA or protein sequences against up-to-date sequence databases.

1.3.2 Sequence alignment provides insights into evolutionary relationships and structure and function of a DNA or protein:

Researchers are often interested to compare two sequences to determine their evolutionary relationships or to judge any similarity in their structure or function. Sequence alignment can be defined as a sequential display of two or more sequences allowing gaps so as to reveal maximum similarity. In other words, sequence alignment calculates minimum number of mutational events required to convert one sequence into another. Sequence alignment would have been easy if there were no insertions or deletions occurring in DNA. In such cases, optimal alignment could be obtained merely by sliding one sequence against another. However, since insertions or deletions occur frequently in genomic DNA, the sequence alignment problem becomes complex and computationally challenging.

Needleman and Wunsch (1970) were the first to propose a rigorous method applicable to find similarity between two sequences. Briefly, to align two sequences of length m and n, an m x n matrix is created where each element mat_{ij}, represents similarity score between i^{th} and j^{h} elements of two sequences. Then, the maximum match pathway is obtained by beginning at the terminals of the sequences (bottom-right corner) and proceeding towards origin (top-left corner). The maximum alignment score at position (i,j) is obtained by adding mat_{ij} score plus the higher value among either of alignment score at $mat_{(i+1)(j+1)}$ or maximum alignment score in row (i+1) or column (j+1) after correcting for gap penalty. At this stage, each of the alignment score represents maximum similarity between subsequences starting at positions (i) and (j) in two sequences. When one reaches the origin, the maximal alignment score is already obtained and then alignment is written by

tracing back the path. Using appropriate scoring matrix and gap penalty, this algorithm guaranties to reveal maximum alignment between two sequences. Smith and Waterman (1981) algorithm is another widely used method that provides an efficient modification of dynamic programming algorithm to identify common molecular subsequences between two long sequences.

Simultaneous alignment of several related sequences is useful to determine phylogenetic relationships and identify structurally or functionally important conserved regions. For example, when Yanofsky et al., (1990) cloned *agamous* homeotic gene from *Arabidopsis*, its deduced amino acid sequence showed similarity to *Antirrhinum* homeotic gene *def A*, human *Serum Response Factor* and yeast transcription factor MCM-1. Multiple alignment of these sequences revealed ~56 residue conserved region (termed MADS Box domain) important for DNA binding and dimerization. Later, this domain has been found to be conserved in a large number of transcription factors involved in key developmental processes in a diverse range of eukaryotes including yeast, plants, insects and mammals (Shore and Sharrocks, 1995).

Clustal-W (Thompson et al., 1994) is one of the most commonly used multiple sequence alignment programs. It first calculates pairwise similarity scores between all possible sequence combinations and then based on similarity matrix a tree is constructed following neighbor joining method. Using such a tree, most similar sequences are progressively aligned substituting the consensus for each pair as they are aligned. Clustal-W also allows differential weighting of sequences based on the extent to which they are related. Users can select various alignment parameters like scoring matrix, k-tuple size, gap-opening and gap-extension penalties.

1.3.3 Understanding phylogenetic relationships using sequence data:

DNA replication may not be always faithful due to rare failure of DNA proof reading machinery or due to misreading of a base that has undergone chemical change. On an average, 1 in 10^6-10^8 nucleotides is likely to be mutated per generation in different

species (Avise, 1994) and the sustenance of such mutations depends on the extent to which they affect the function. If two evolutionarily related sequences having a common ancestor are aligned, the changes that have occurred in them can be revealed. Since mutation rates at a particular locus are fairly constant over time (molecular clock), it is safe to assume that higher the number of changes observed in two orthologous sequences longer is the evolutionary time elapsed since their divergence. Thus, multiple sequence alignment of related sequences can be used to construct evolutionary history of genes or of the species to which they belong.

For construction of phylogenetic tree from sequence data, either distance matrix methods or character set methods are used (Saitou, 1996). In distance matrix methods, first, all the sequences are aligned in pairwise combinations and genetic distances between them are calculated. Then using distance matrix, trees are constructed following UPGMA (unweighted pair group method using arithmetic mean) or Neighbor Joining Method. In the programs that use complete character data set (e.g. multiple alignment), information at all orthologous positions in a set of sequences is considered while building a tree. Minimum Evolution Method, Maximum Parsimony Method and Maximum Likelihood Method are some of the important approaches that use complete character data for phylogenetic analysis. Researchers have extensively used nucleotide and/or amino acid sequence information from various genes and/or proteins like globins, actins, histones, cytochromoes, Rubisco, ribosomal-DNA, mitochondrial DNA and/or chloroplast DNA to decipher evolutionary relationships between diverse families, genera, species and populations.

Though sequence data is highly useful to understand relative evolutionary distance, care is essential before estimating time of divergence assuming constant molecular clock. This is mainly because, different nucleotide / amino acid sites in a sequence can evolve at markedly different rates due to their varying functional significance and varying selection pressure. Indeed, it is known that mutation rates vary considerably depending on various factors such as the position of a nucleotide within a codon (synonymous / non-synonymous substitution), region of a gene (exon or intron), function of a gene and

region of a DNA (coding / non-coding) (Avise, 1994). Besides, parallel mutational events and the reversible nature of substitutions may confuse the estimated number of mutations (molecular ticks) and thus estimated divergence time. Nonetheless, with appropriate care, DNA and protein sequences provide one of the finest records of evolutionary history ever since the life originated on the earth.

1.3.4 Prediction of RNA secondary structure:

Ribonucleic acids play important role in the flow of information from DNA to protein. The messenger RNAs (mRNA) act as template for protein synthesis whereas ribosomal RNAs (rRNA) and transfer RNAs (tRNA) form integral components of the protein synthesis machinery. Unlike DNA, ribonucleic acids are synthesized as single stranded molecules with every nucleotide having potential to base pair with every other complementary nucleotide in the same molecule. The major stabilizing events in RNA folding comprise spontaneous emergence of double stranded stretches resulting from base pairing across complementary regions. These helical regions are generally short (<10 bp) and often interrupted by bulges, loops, hairpins, pseudo-knots or junctions. In contrast to DNA, double stranded RNA helices adopt A-form structures and in addition to Watson-Crick base pairing (A-U and G-C), they show several non-canonical base pairs like G-U, U-C, U-U and highly stable tetraloops like 5'UNCG-3' or 5'GNRA-3' (Puglisi and Puglisi, 1997). An optimal three-dimensional folding of an RNA molecule minimizes free energy of the solution facilitated by maximal base pairing between the complementary nucleotides and plays important role in stability, catalytic function, splicing and/or localization of RNA.

Several methods have been proposed for the prediction of RNA secondary structure using the nucleotide sequence information of an RNA molecule which helps in understanding RNA function (e.g. Tinco et al., 1971; Zuker and Stiegler, 1981; Gautheret et al., 1990; Zuker, 2000). These methods consider the experimentally calculated free energy parameters for RNA base pairing, helix initiation, helix progression and hairpin formation (e.g. Freier et al., 1986). Then, from the vast search space involving all

possible combinations, most optimal folding stretches are calculated employing dynamic programming algorithm. In addition, information about the experimentally identified double stranded regions or phylogenetically conserved regions can be used to narrow down the search space (Jaeger et al., 1990; Gorodkin et al., 1997). Availability of powerful computing machines and efficient algorithms now permits reliable RNA secondary structure prediction of not only the small molecules like tRNA, but even the large viral RNA genomes like HIV and Hepatitis C Virus (Hofacker et al., 1998).

1.3.5 Predicting novel genes in genome sequences:

In high-throughput genome sequencing projects, once large genome sequences are assembled the next immediate step involves masking of repetitive/low complexity regions followed by identification of genes contained in it. Though a database search can conveniently detect previously characterized genes, the real computational challenge lies in the prediction of novel genes in genome sequences. Since a large number of genes are unknown in different organisms, prediction of new genes can help researchers to design experiments to determine their function and discover missing links in metabolic pathways or other life processes.

Predicting genes in bacterial genomes is relatively easy since bacterial protein coding regions consist of contiguous Open Reading Frames (Claverie, 1997). On the other hand, identification of novel genes in eukaryotic genome sequences is a more complex problem since eukaryotic protein coding regions are frequently interrupted by non-coding introns. For example, the human genome sequencing has revealed that a single human gene may contain upto 178 exons with an average of ~9 exons per gene (International Human Genome Sequencing Consortium, 2001). Moreover, for human genes, the length of exons may range from ~19 bp to more than ~17 kbp (mean ~145 bp, median ~122 bp) while an intron may span from ~60 bp to more than ~30 kbp (mean ~3300 bp, median ~1023 bp). Overall, only ~1.1% of the human genome sequence has been found to be encoded by exons and ~24% by introns with remaining ~75% of the sequence is intergenic (Venter et

al., 2001). The intricacy of finding genes in genome sequences is further complicated by the possibility of overlapping genes and alternate splicing sites.

Prediction of genes in eukaryotic genome sequences is essentially the accurate prediction of exons and then constructing complete gene structures. For the ease of prediction, 4 types of exons can be distinguished: initial exon (transcription initiation site to first 5' splice junction), internal exon (3' splice junction to 5' splice junction), terminal exon (3' splice site to transcription termination signal) and single exon genes (Burge and Karlin, 1998). The sequence features that help in prediction of exons include transcription signals, translation signals, splicing signals and characteristic frequencies of nucleotides in coding and non-coding regions. Moreover, the similarity of a predicted exon to a known EST, cDNA or gene, not only confirms the reliability of prediction but also helps in assigning function. After finding out the most probable exons, the next important task involves the prediction of a complete gene structure so that they can be fitted into a coherent gene model. A complete gene structure would constitute promoter regions with putative regulatory elements, the transcription initiation site, clearly marked introns and exons, translation signals, protein coding region, and poly-adenylation signals.

Several computational tools have been developed that predict putative exons and complete gene models in genome sequences. The GENESCAN program (Burge and Karlin, 1997) uses a general probabilistic model of gene structure incorporating basic transcriptional, translational and splicing signals as well as length distributions and compositional features of exons, introns and intergenic regions. The GRAIL suite of programs (Uberbacher et al., 1996) use neural network method for prediction of protein coding exons and polymerase-II promoter regions in genome sequences. In the *Genie* gene finding program, hidden Markov models are used integrating the information based on signal sensors (e.g. splice sites, start codon), content sensors (exons, introns and intergenic regions) and alignments with mRNA, ESTs and peptide sequences (Reese et al., 2000). It has been observed that configuration of gene prediction programs with species specific content / signal parameters greatly help in accurate prediction of novel genes in genome sequences.

1.3.6 Protein sequence motif analysis helps in understanding protein function:

Many proteins are localized in the specific compartments like mitochondria, chloroplast, nucleus or membrane of the cell. This information for protein targeting is generally encoded in the N-terminal domain of the protein spanning ~20-30 amino acid residues. Signal peptide sequences and their cleavage sites have been studied from a large number of proteins and this knowledge base helps in the prediction of protein targeting signal and subcellular localization of new proteins (Claros et al., 1997).

Protein structures are frequently modular in nature with distinct domains performing distinct functions. In a protein, often, a few specific regions play important role in protein function, for example in their binding properties or enzyme activity, and tend to be conserved in both structure and sequence (Hofmann et al., 1999). These conserved regions are known as motifs, patterns, profiles or signatures and can be identified from multiple alignment of related protein sequences. Protein sequence motifs may span from a few to several residues and the degree of conservation may vary across the motifs. Identification and characterization of protein sequence motifs or critical residues in them, is important in understanding protein function.

For comprehensive description of a motif, it is essential that diverse members of a protein family are selected and an accurate alignment of the sequences is obtained. However, care should be taken not to over-represent too similar sequences and apart from sequence conservation, there should be additional experimental evidence to demonstrate that the conserved region is important for structure or function of the protein (Bork and Gibson, 1996). From the alignment block, motif sequence can be described as a consensus sequence or as a weight matrix description. In the consensus method, all the observed residues at a particular location are considered equally likely. However, this is rarely a case and different residues at a position can have different specificities. The profile or position weight matrix (PWM) of a motif considers the relative frequency of observed residues as an indicator of their relative importance. Thus, weight matrices are more sensitive and provide quantitative estimate of motif function.

Searching a query sequence against a collection of known sequence motifs helps to identify significant pattern(s) that permits assignment of function to new proteins. Similarly, analysis of a protein sequence database using a motif signature allows detection of new members of a protein super family. The PROSITE database (Hofmann et al., 1999) is one of the earliest attempts in compilation of protein sequence motifs initiated since 1988. In PROSITE, protein sequence motifs have been originally defined as consensus sequences. However, considering the specificity and sensitivity offered by weight matrices, efforts are being made to define them as profile descriptions. The BLOCKS database (Henikoff et al., 2000) provides a collection of ungapped multiple alignments of motif sequences of diverse members of various protein families. The Pfam database describes modular domain structures of protein families using profile hidden Markov models (Bateman et al., 2000). SMART (Simple Modular Architecture Research Tool) is another useful database that allows identification and annotation of genetically mobile domains and domain architecture (Schultz et al., 2000).

1.3.7 Prediction of protein structure from its amino acid sequence:

The major determinant of the protein structure is the peptide backbone itself with its characteristic planar peptide bond, phi-psi rotations and dual hydrogen bonding capabilities. With repeating values of certain phi-psi angles, the hydrogen bond donors and acceptors can be juxtaposed and hydrogen bonding can be satisfied within the peptide backbone itself. Two types of repeating structures are common in proteins: alpha-helices and beta-sheets. The alpha-helix is formed by repeated hydrogen-bonds between carboxyl group of $(n)^{th}$ residue and amino group of $(n+4)^{th}$ residue, with repeated phi-psi values of about -60^0 and -40^0. On the other hand, the beta-strands are formed by repeated phi-psi values near -120^0 and 140^0 and regular hydrogen-bonds extending from one strand to another similarly aligned strand (Richardson and Richardson, 1989). Often, during protein folding, these secondary structures arise spontaneously by local interactions. Further, adjacent secondary units pack and serve as building blocks, folding nuclei or domains and consequently, the correct arrangement of all functional groups in space facilitates the protein function.

In 1961, Anfinsen and coworkers showed that ribonuclease could be completely denatured and then refolded again without loss of enzymatic activity implying that amino acid sequence of a protein contains sufficient information for its correct 3-dimensional confirmation in a particular environment. This observation prompted researchers to develop methods to predict the structure of a protein from its sequence alone.

Though secondary structures are stabilized by peptide backbone groups, the side chains of amino acids can have profound effect on secondary structure by imparting constraints on rotation and by competing for hydrogen-bonds. Analyses of several high-resolution protein structures have helped to determine how a particular residue behaves in a particular confirmation. For example, asparagine and serine are frequently found at helix initiation (N-cap) whereas glycine is one of the most common helix terminator (C-cap) (Richardson and Richardson, 1989). On the other hand, conformational constraints of proline make it unfit for any regular secondary structure and many a times it acts as a breaker of alpha helix and beta-strand. The protein secondary structure prediction programs make use of this knowledge base and calculate propensity of the given polypeptide region to adopt a particular secondary structure.

For the prediction of protein secondary structure, Chou-Fasman method uses the propensities of amino acid residues for occurrence in helix, beta-sheet and coil, calculated from datasets of high resolution structures (Prevelige and Fasman, 1989). The GOR method which uses information theory for secondary structure prediction (Granier and Robson, 1989), assumes that confirmation of a residue is not only determined by the nature of the residue itself but by every other residue in the neighborhood. However, for the ease of prediction, GOR method considers effects of 8 residues on each side of a residue. The PHD method uses neural-network approach for prediction of secondary structure (Rost, 1996). Whenever a new sequence is submitted, it is scanned against a database and the profile multiple alignment obtained is fed through the pre-trained neural network, which gives probability of each residue to be in helix, strand or loop.

Prediction of protein tertiary structure is a more complex problem unless a sequence homolog is available in the structure database. In homology modeling, first the query sequence is scanned against the structure database to pick up structures with a similar sequence. A reference structure with more than ~30% identity spanning along a major part of the sequence is considered highly useful. The backbone co-ordinates of the reference structure are obtained and residues from the query sequence are threaded at analogous positions. Energy-minimization steps are then carried out so that all the insertions-deletions can be smoothened and all the residues adopt optimal positions under given constraints. In the absence of structure homologs, sequence structure threading or *ab initio* methods can be attempted.

1.4 Genesis of Thesis:

When I joined the Plant Molecular Biology Unit at National Chemical Laboratory, my initial work was on identification of molecular markers linked to some quality characters in wheat. During the same time, I started learning computer programming using a "C" compiler installed on a computer in the laboratory. I found the data encoding and data handling tools in "C" to be very efficient and realized that with the logical design and implementation of programs, computers could be easily harnessed for large-scale computation.

While learning computer programming I always used to think how it could be applied to solve biological problems. I realized that availability of a large amount of sequence data provides enormous opportunities to unravel biological information hidden in DNA and protein sequences. This thesis is thus an outcome of my efforts to apply computer programming for analysis of biomolecular sequences to address some specific questions that occurred to me.

The specific objectives of my thesis were:

1. To assess the organization of simple sequence repeats in eukaryotic genome sequences.
2. To analyze the extent of codon reiterations in complete genome coding DNA sequences of yeast, *C. elegans* and *Drosophila*.
3. To study occurrence of various kinds of internal repeats in protein sequences and analyze implications of repetitive sequence patterns on protein structure and function.
4. To systematically describe and develop a database of Tandem Repeats in Protein Sequences.
5. To develop computational tools for comparative promoter sequence analysis.

The results of the work carried out to fulfil the above objectives have been presented in the form of following four chapters apart from the present chapter of introduction:

Chapter 2 : Differential distribution of simple sequence repeats in eukaryotic genome sequences

Chapter 3 : Amino acid repeat patterns in protein sequences : Their diversity and structural-functional implications

Chapter 4 : Development of a web based software tool, *TRES*, for comparative promoter sequence analysis

Chapter 5 : Thesis overview

References

ABSTRACT:

Simple sequence repeats are ubiquitous in eukaryotic genome sequences and are thought to contribute in genome organization and evolution. I analyzed complete chromosome sequences available from human, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* (yeast) for the occurrence of mono-, di-, tri- and tetranucleotide repeats. All these genomes exhibited characteristic microsatellite distributions. I observed that though the trends for different repeats were similar between different chromosomes within a genome, the density of repeats might vary between different chromosomes of the same species. Abundance or rarity of various di- and trinucleotide repeats in different genomes could not be explained by nucleotide composition of a sequence or potential of repeated motifs to form alternative DNA structures. This suggests that in addition to nucleotide composition of repeat motifs, characteristic DNA replication/repair/recombination machinery might play an important role in the genesis of repeats. Moreover, analysis of complete genome coding DNA sequences of *Drosophila*, *C. elegans* and yeast have indicated that expansions of codon repeats corresponding to small hydrophilic amino acids are tolerated more, while strong selection pressures probably eliminate codon repeats encoding hydrophobic and basic amino acids. The locations and sequences of all the repeat loci detected in genome sequences and coding DNA sequences have been made available at the URL: http://www.ncl-india.org/ssr and could be useful in further genetic studies.

2.1 INTRODUCTION:

Mutation, recombination and duplication are key events that generate variability and thereby facilitate evolution. DNA duplications occur at various levels including entire genome (polyploidy), chromosome, part of a chromosome, entire gene or DNA segments of a few to hundreds of bases. Among various DNA duplication events, simple sequence repeats (SSR) or microsatellites are the genetic loci where one or a few bases are tandemly repeated for a varying number of times. Simple sequence repeats in DNA originate primarily due to slipped-strand mis-pairing and subsequent error(s) during DNA replication, repair or recombination (Levinson and Gutman, 1987; Figure 2.1). Microsatellite loci mutate by insertions or deletions of one or a few repeat units and with the increase in the length of repeat tracks, these loci become more susceptible to DNA strand slippage and show elevated mutation rates (Wierdl et al., 1997). It has been observed that microsatellite loci are highly unstable with the mutation rates of $\sim 10^{-3}$ to 10^{-6} per locus per generation as compared to normal DNA base substitution rates of $\sim 10^{-6}$ to 10^{-9} per base per generation (Ellegren, 2000b). Consequently, microsatellite loci show high length polymorphism and therefore, they are widely used in DNA-fingerprinting and diversity studies. Moreover, since they can be easily assayed by PCR using unique flanking primers, they are considered to be ideal genetic markers for construction of high density linkage maps (Beckmann and Soller, 1990; Morgante and Olivieri, 1993).

Simple sequence repeats are densely interspersed in eukaryotic genomes and are thought to be a major source of quantitative genetic variation (Kashi et al., 1997) and have been also implicated in promoting recombination (Majewski and Ott, 2000). During the past decade, several human neurodegenerative diseases have been found to be associated with dynamic mutations occurring at microsatellite loci within or near to specific genes (Ashley and Warren, 1995; Table 2.1). Therefore, there has been an increased interest to understand the molecular mechanisms involved in origin, evolution and expansion / deletion of microsatellites. Formation of alternative DNA structures, like hairpins or intramolecular triplexes, have been implicated in expansions of $(CTG)_n.(CAG)_n$, $(CCG)_n.(CGG)_n$ and $(GAA)_n.(TTC)_n$ repeats (Pearson and Sinden, 1998; Mitas, 1997).

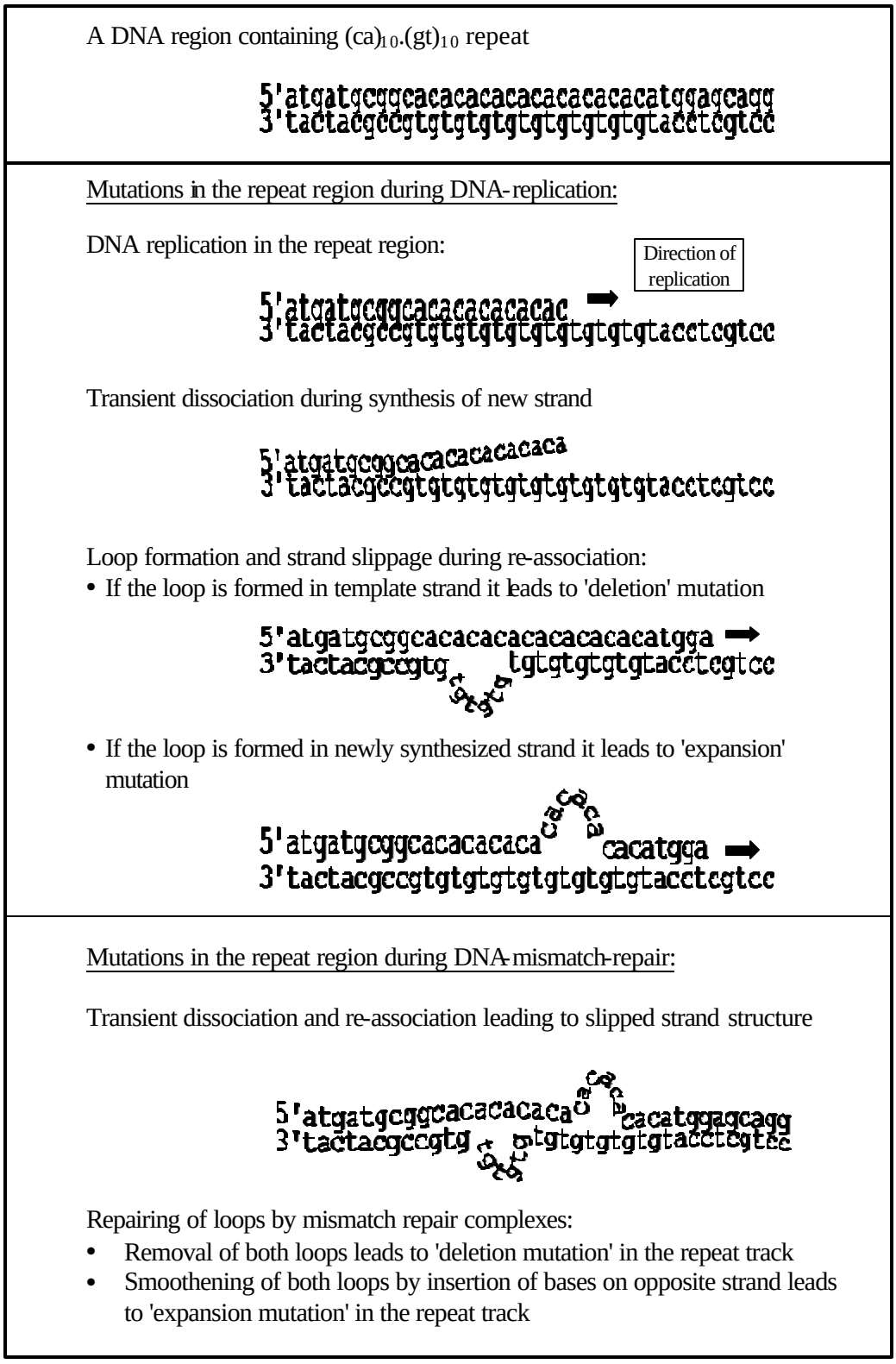


Figure 2.1 A schematic diagram showing mechanism of strand slippage involved in the genesis of simple sequence repeats

Table 2.1: Trinucleotide repeats in human genetic diseases [a]

Disease	Repeat	Repeat length		Possible biological effect of expansion
		Normal	Pathogenic	
Fragile XA	$(CGG)_n$	6-52	230-2000	Promoter methylation, gene silencing
Fragile XE	$(CCG)_n$	4-39	200-900	- „ -
Fragile XF	$(CGG)_n$	7-40	306-1008	- „ -
Fragile 16a	$(CCG)_n$	16-49	1000-1900	- „ -
Jacobesen syndrome (FRA11B)	$(CGG)_n$	11	100-1000	- „ -
Kennedy syndrome (SMBA)	$(CAG)_n$	14-32	40-55	Polyglutamine expansion, protein malfunctioning
Huntington disease (HD)	$(CAG)_n$	10-34	40-121	- „ -
Spinocerebellar ataxia 1 (SCA1)	$(CAG)_n$	6-39	40-81	- „ -
Spinocerebellar ataxia 2 (SCA2)	$(CAG)_n$	14-31	34-59	- „ -
Spinocerebellar ataxia 3 (SCA3) / Machado-Joseph disease (MJD)	$(CAG)_n$	13-44	60-84	- „ -
Spinocerebellar ataxia 6 (SCA6)	$(CAG)_n$	4-18	21-28	- „ -
Spinocerebellar ataxia 7 (SCA7)	$(CAG)_n$	7-17	38-130	- „ -
Haw River Syndrome (DRPLA)	$(CAG)_n$	7-25	49-75	- „ -
Mytonic dystrophy (DM)	$(CTG)_n$	5-37	80-1000	Repeats track in 3' UTR , altered mRNA processing
Friedreich ataxia (FRDA)	$(GAA)_n$	6-29	200-900	Repeats track in intron, altered mRNA production

[a] Adapted from Sinden (1999)

Studies in *E. coli* (Kang et al., 1995) and yeast (Freudenreich et al., 1997) have shown that stability of $(CAG)_n$ repeats varies with their orientation relative to direction of replication. This has been attributed to differences in hairpin propensity for $(CAG)_n$ and $(CTG)_n$ stretches that can occur either on leading strand or lagging strand depending on the direction of replication. Microsatellite instability has been also found to be influenced by mutations in genes involved in mismatch repair (Sia et al., 1997) and DNA replication (Kokoska et al., 1998). However, Miret et al., (1997) have observed little or no difference in $(CAG)_n$ repeat instability in yeast strains containing disruptions in mismatch repair genes *MSH2, MSH3* or *PMS1* or recombination gene *RAD52.*

Frequencies of various microsatellite sequences in different genomes have been estimated experimentally by hybridization technique (e.g. Hamada et al., 1982; Tautz and Renz, 1984; Panaud et al., 1995). However, this could not be done accurately using oligo-nucleotides like $(AT)_n$, $(GC)_n$ that can self-complement. With the growth of sequence databases, several authors have reported the abundance of simple sequence repeats in different genomes (e.g. Beckman and Weber, 1992; Wang et al., 1994; Jurka and Pethiyagoda, 1995; Hancock, 1995; Richard and Dujon, 1996; Bachtrog et al., 1999; Kruglyak et al., 2000). In a recent survey, Toth et al., (2000) have examined the distribution of microsatellites in exonic, intronic and intergenic regions of several eukaryotic taxa. Differential abundance of repeats in different genomes has led them to suggest that strand-slippage theories alone are insufficient to explain characteristic microsatellite distributions.

Knowledge of relative distribution of simple sequence repeats in a genome, is essential for understanding mechanisms involved in their genesis. However, most of the previous studies on microsatellite distributions were based on DNA sequence databases over-represented by coding or gene-rich regions. On the other hand, availability of complete genome sequences now permits the determination of frequencies of SSRs at the whole genome level, which should reflect basal level of SSR dynamics within a species. In this chapter, I have analyzed the occurrences of simple sequence repeats in a few eukaryotic genomes where complete genome/chromosome sequences were available. Moreover, I have also studied non-redundant complete genome coding DNA sequences of *Drosophila*, *C. elegans* and yeast to assess the extent of codon reiterations in protein coding regions.

2.2 MATERIALS AND METHODS:

All the genome sequences were downloaded in FASTA format from ftp://ncbi.nlm.nih.gov/genbank/genomes/. The list of genome sequences and their lengths are shown in Table 2.3. The human chromosome-21 (Hattori et al., 2000) and -22 (Dunham et al., 1999) sequences were obtained as ensemble of 5 and 12 contig sequences, respectively. Individual chromosome sequences of *Saccharomyces cerevisiae* (Goffeau et al., 1996), *Caenorhabditis elegans* (The *C. elegans* sequencing

consortium, 1998) and *Arabidopsis thaliana* chromosome-II (Lin et al., 1999) and -IV (Mayer et al., 1999) were available as single contiguous strings. The *C. elegans* chromosome sequences have a few unsequenced gaps represented as stretches of "N" in the sequences and the lengths shown in Table 2.3 are corrected by removing such gaps. Most of the *Drosophila melanogaster* genome has been sequenced by whole-genome shotgun sequencing (Adams et al., 2000) and sequences have been made available as a collection of scaffolds. Only the genomic scaffolds mapped on chromosome-X, 2 and 3 were selected and obtained using GenBank's Batch Entrez facility. Accession numbers or links to all the sequences used in this study are available at http://www.ncl-india.org/ssr.

All the genome sequences were scanned for various SSRs using computer programs written in "C". A simple sliding window technique was used for detection of tandem repeats. Figure 2.2 depicts a representative flowchart of the algorithm employed for searching simple sequence repeats in large genome sequences. Briefly, consider a DNA sequence as a string, $B_1B_2B_3B_4B_5$……B_i……$B_{n-1}B_n$. To detect a tandem repeat of size (k=1 to 4) at position (i), the window "B_i…B_{i+k-1}" was compared with subsequent windows starting at positions B_{i+k}, B_{i+2k}, B_{i+3k}, B_{i+4k}….. (k=1 for mono-, k=2 for di-, k=3 for tri- and k= 4 for tetranucleotide repeats). A repeat was detected and extended further when a certain minimum number of units (20, 10, 7 or 5 for mono-, di-, tri-, or tetranucleotide repeats, respectively) were repeated tandemly. Repeats were searched allowing a maximum of one mismatch for every 10 nucleotides. While scanning for di-, tri- and tetranucleotide repeats, combinations involving the runs of same nucleotide were not considered. Similarly, for tetranucleotide repeats, combinations representing perfect dinucleotide repeats were ignored. Significance of difference in density of repeats between different chromosomes of the same species was determined using '*t* test'. Frequency distributions of repeats along 1 mbp (million base pairs) contiguous segments of a chromosome were used for calculation of variance for '*t* test'. However, the significance could be tested only for human, *Arabidopsis* and *C. elegans* sequences, where long contiguous chromosome sequences were available.

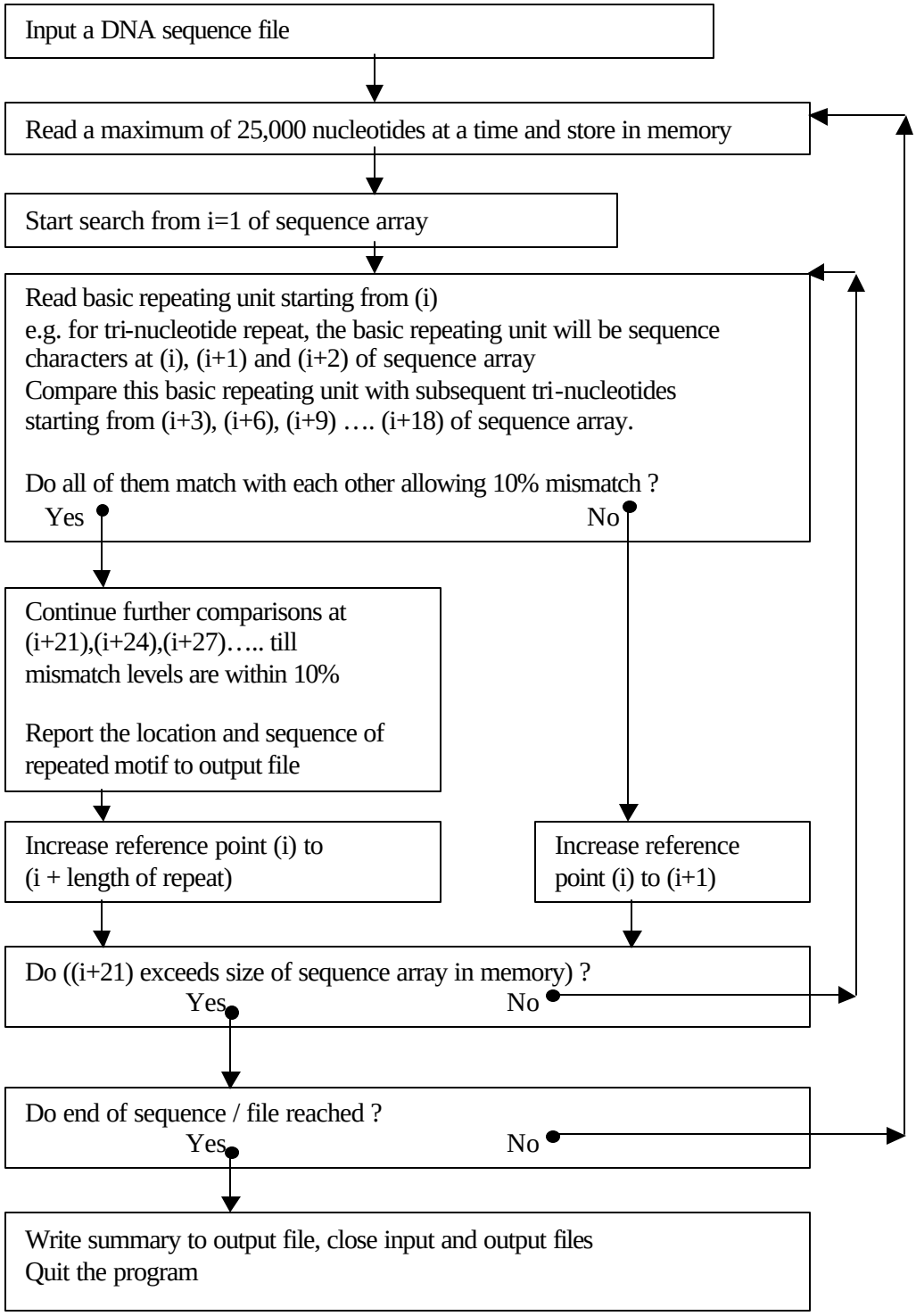


Figure 2.2: Flowchart showing the algorithm used for searching trinucleotide repeats in large genomic DNA sequences.

Table 2.2: Descriptions of various types of mono-, di-, tri- and tetranucleotide repeat classes

Representative sequence of repeat class	Equivalent combinations on same strand in different reading frames				Equivalent combinations on complementary strand in different reading frames			
Mononucleotide repeats								
A	A				T			
G	G				C			
Dinucleotide repeats								
AT*	AT		TA					
AG	AG		GA		CT		TC	
AC	AC		CA		GT		TG	
GC*	GC		CG					
Trinucleotide repeats								
AAT	AAT	ATA	TAA		ATT	TTA	TAT	
AAG	AAG	AGA	GAA		CTT	TTC	TCT	
AAC	AAC	ACA	CAA		GTT	TTG	TGT	
ATG	ATG	TGA	GAT		CAT	ATC	TCA	
AGT	AGT	GTA	TAG		ACT	CTA	TAC	
AGG	AGG	GGA	GAG		CCT	CTC	TCC	
AGC	AGC	GCA	CAG		GCT	CTG	TGC	
ACG	ACG	CGA	GAC		CGT	GTC	TCG	
ACC	ACC	CCA	CAC		GGT	GTG	TGG	
GGC	GGC	GCG	CGG		GCC	CCG	CGC	
Tetranucleotide repeats								
AAAT	AAAT	AATA	ATAA	TAAA	ATTT	TTTA	TTAT	TATT
AAAG	AAAG	AAGA	AGAA	GAAA	CTTT	TTTC	TTCT	TCTT
AAAC	AAAC	AACA	ACAA	CAAA	GTTT	TTTG	TTGT	TGTT
AATT*	AATT	ATTA	TTAA	TAAT				
AATG	AATG	ATGA	TGAA	GAAT	CATT	ATTC	TTCA	TCAT
AATC	AATC	ATCA	TCAA	CAAT	GATT	ATTG	TTGA	TGAT
AAGT	AAGT	AGTA	GTAA	TAAG	ACTT	CTTA	TTAC	TACT
AAGG	AAGG	AGGA	GGAA	GAAG	CCTT	CTTC	TTCC	TCCT
AAGC	AAGC	AGCA	GCAA	CAAG	GCTT	CTTG	TTGC	TGCT
AACT	AACT	ACTA	CTAA	TAAC	AGTT	GTTA	TTAG	TAGT
AACG	AACG	ACGA	CGAA	GAAC	CGTT	GTTC	TTCG	TCGT
AACC	AACC	ACCA	CCAA	CAAC	GGTT	GTTG	TTGG	TGGT
ATAG	ATAG	TAGA	AGAT	GATA	CTAT	TATC	ATCT	TCTA
ATAC	ATAC	TACA	ACAT	CATA	GTAT	TATG	ATGT	TGTA
ATGG	ATGG	TGGA	GGAT	GATG	CCAT	CATC	ATCC	TCCA
ATGC*	ATGC	TGCA	GCAT	CATG				
ATCG*	ATCG	TCGA	CGAT	GATC				
AGAC	AGAC	GACA	ACAG	CAGA	GTCT	TCTG	CTGT	TGTC
AGTG	AGTG	GTGA	TGAG	GAGT	CACT	ACTC	CTCA	TCAC
AGTC	AGTC	GTCA	TCAG	CAGT	GACT	ACTG	CTGA	TGAC
AGGT	AGGT	GGTA	GTAG	TAGG	ACCT	CCTA	CTAC	TACC
AGGG	AGGG	GGGA	GGAG	GAGG	CCCT	CCTC	CTCC	TCCC
AGGC	AGGC	GGCA	GCAG	CAGG	GCCT	CCTG	CTGC	TGCC
AGCT*	AGCT	GCTA	CTAG	TAGC				
AGCG	AGCG	GCGA	CGAG	GAGC	CGCT	GCTC	CTCG	TCGC
AGCC	AGCC	GCCA	CCAG	CAGC	GGCT	GCTG	CTGG	TGGC
ACGT*	ACGT	CGTA	GTAC	TACG				
ACGG	ACGG	CGGA	GGAC	GACG	CCGT	CGTC	GTCC	TCCG
ACGC	ACGC	CGCA	GCAC	CACG	GCGT	CGTG	GTGC	TGCG
ACCG	ACCG	CCGA	CGAC	GACC	CGGT	GGTC	GTCG	TCGG
ACCC	ACCC	CCCA	CCAC	CACC	GGGT	GGTG	GTGG	TGGG
GGGC	GGGC	GGCG	GCGG	CGGG	GCCC	CCCG	CCGC	CGCC
GGCC*	GGCC	GCCG	CCGG	CGGC				

*Nucleotide combinations in these repeat classes are self-complementary.

A poly-(A) repeat is same as poly-(T) repeat on a complementary strand. Similarly, $(AC)_n$ is equivalent to $(CA)_n$, $(TG)_n$ and $(GT)_n$ while, $(AGC)_n$ is equivalent to $(GCA)_n$, $(CAG)_n$, $(CTG)_n$, $(TGC)_n$ and $(GCT)_n$ in different reading frames or on a complementary strand. Thus, two unique classes are possible for mononucleotide repeats, whereas four classes are possible for di-, ten for tri- and thirty-three for tetranucleotide repeats (Table 2.2). I have determined individual repeat frequencies for all these classes.

Complete genome coding DNA sequences of all predicted peptides of *Drosophila*, *C. elegans* and yeast were obtained from the Berkeley Drosophila Genome Project (http://www.fruitfly.org), the Sanger Centre's Wormpep Database (http://www.sanger.ac.uk/Projects/C_elegans/wormpep) and the *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces), respectively. A codon repeat was considered only when it was tandemly repeated for a minimum of 7 times allowing 1 mismatch for every 10 nucleotides.

All computer programs were implemented on a personal computer with a Pentium Pro(R) microprocessor and 16 MB RAM. A typical program took less than ~5 minutes to analyze 10 mbp of sequence. Outputs of the programs were verified by comparing some of the repeat loci in the original sequence. The results were compiled and this resource has been made available at http://www.ncl-india.org/ssr

2.3 RESULTS AND DISCUSSION:

While searching a sequence for simple sequence repeats, defining the minimum number of repeats and mismatch considerations are important empirical criteria. For detection of various repeats in genome sequences, we selected minimum repeating units such that a repeat spans for a minimum of 20 nucleotides. Although previous studies have used threshold repeat lengths of 10-12 nucleotides, any preference(s) in genesis of repeats or variations in mutation rates are likely to be more clear at longer threshold lengths. Besides, longer repeats being more unstable, have implications in genome organization, genetic variation, protein evolution and disease, at a relatively shorter evolutionary time scale. Simple sequences can be pure tandem repeats or may contain interruption(s) due to accumulation of point mutation(s) or can have

scrambled arrangement of repetitive motifs (Tautz et al., 1986). However, most of the previous studies have considered only perfect repeats without allowing any mismatch. I observed that several long repeats contain one or a few base substitutions and hence, if only perfect repeats are considered, such loci are likely to be counted as two or more separate repeats of shorter lengths. Therefore, rather than considering only perfect repeats, I allowed one mismatch for every 10 nucleotides. Although appearance of mismatches in repeats can reduce the chances of slippage-mediated expansions / deletions (Petes et al., 1997), such loci might represent previous occurrences of perfect repeats. Moreover, interruption(s) in a repeat track may be only a transition state and could be removed by DNA replication slippage or reverse mutation(s) (Harr et al., 2000).

2.3.1 Characteristic trends in microsatellite distributions:

Analysis of complete genome/chromosome sequences, available from human, *Drosophila*, *Arabidopsis*, *C. elegans* and yeast (Table 2.3, Figure 2.3), has revealed that compared to other genomes, human chromosomes 21 and 22 are rich in mono- and tetranucleotide repeats. On the other hand, the *Drosophila* chromosomes have higher frequency of di- and trinucleotide repeats. Surprisingly, *C. elegans* genome contains less number of SSRs per mbp of sequence compared to yeast genome. Moreover, the frequency of trinucleotide repeats in yeast is more than that observed in human chromosomes-21 and -22.

In all the genomes, among mononucleotide repeats, poly-(A) / poly-(T) repeats were predominant while poly-(C) / poly-(G) repeats were rare. Tetranucleotide repeats were highly frequent in human chromosomes and most common among them were $(AAAT)_n$, $(AAAG)_n$, $(AAAC)_n$, $(ATAG)_n$, $(AAGG)_n$, $(ATGG)_n$ and $(AGGG)_n$. The *Drosophila* chromosomes also contained a large number of tetranucleotide repeats of which $(ATAC)_n$, $(AAAT)_n$, $(AAAC)_n$, $(AGTC)_n$ and $(AACC)_n$ were more frequent. Overall, tetranucleotide repeats of type $(AAAN)_n$ seemed to be more common compared to other combinations.

Table 2.3: Frequency of repeat loci per mbp of individual chromosome sequences in different eukaryotic genomes

Chromosome / arm	Sequence length mbp	Frequency of repeats >= 20 nucleotides				Frequency of repeats >= 40 nucleotides			
		Mono-nucleotide repeats	Di-nucleotide repeats	Tri-nucleotide repeats	Tetra-nucleotide repeats	Mono-nucleotide repeats	Di-nucleotide repeats	Tri-nucleotide repeats	Tetra-nucleotide repeats
Human									
Hs-21	33.82	141.8	105.0	24.8	119.7	3.7	21.3	2.4	15.1
Hs-22	33.62	223.4	81.0	39.0	151.5	4.8	17.4	2.9	17.3
Drosophila									
Dm-X	21.95	157.0	215.1	135.8	96.8	0.8	9.5	7.3	4.2
Dm-2L	22.58	47.5	94.6	62.3	51.9	0.2	2.1	1.8	1.0
Dm-2R	21.07	45.4	102.7	79.0	57.4	0.3	3.3	2.9	1.7
Dm-3L	23.67	56.2	92.3	83.0	55.4	0.3	2.2	2.7	1.2
Dm-3R	27.86	53.8	104.9	85.0	58.0	0.3	3.5	2.5	1.5
Arabidopsis									
At-2	19.65	53.5	51.1	44.2	18.8	0.7	7.8	1.37	0.1
At-4	17.55	53.6	53.6	48.0	17.7	0.5	6.8	1.48	0.2
C. elegans									
Ce- I	14.75	37.5	34.8	28.8	21.2	0.1	4.7	0.61	0.6
Ce-II	16.62	30.4	22.4	25.8	25.3	0.1	3.1	0.60	0.4
Ce-III	11.60	30.3	30.9	31.8	19.4	0.0	3.7	0.43	0.3
Ce-IV	14.45	23.2	22.0	23.9	23.9	0.0	2.1	0.21	0.5
Ce-V	20.52	27.6	17.4	18.1	18.4	0.1	2.9	0.24	1.1
Ce-X	17.29	30.8	30.0	20.2	15.3	0.2	4.1	0.40	0.2
Yeast, all 16 chromosomes	12.07	44.2	31.7	50.0	12.3	1.8	2.4	4.89	0.3

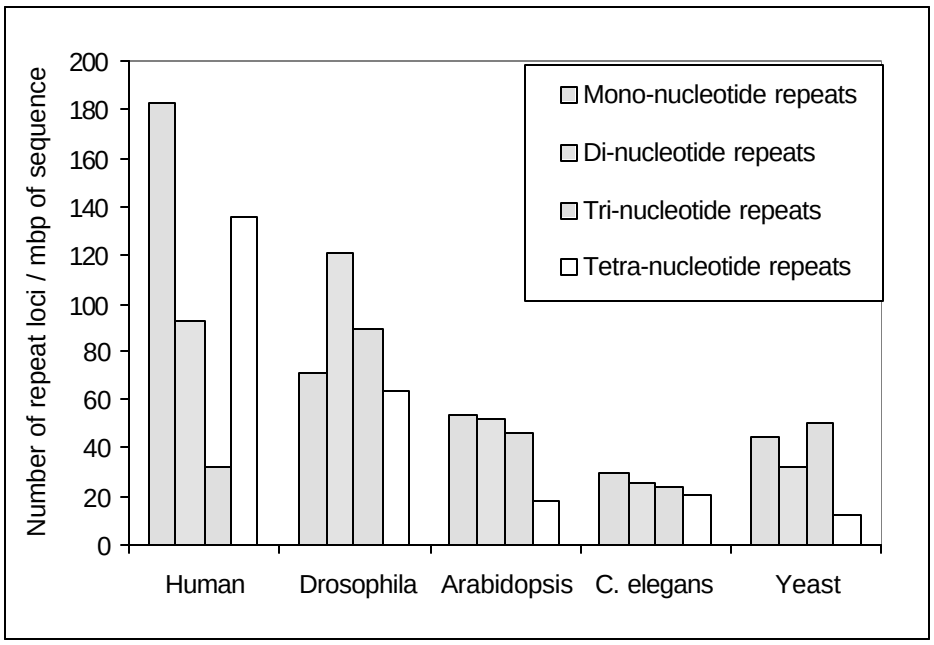


Figure 2.3 Frequency of repeat loci per mbp of chromosome sequences in different genomes

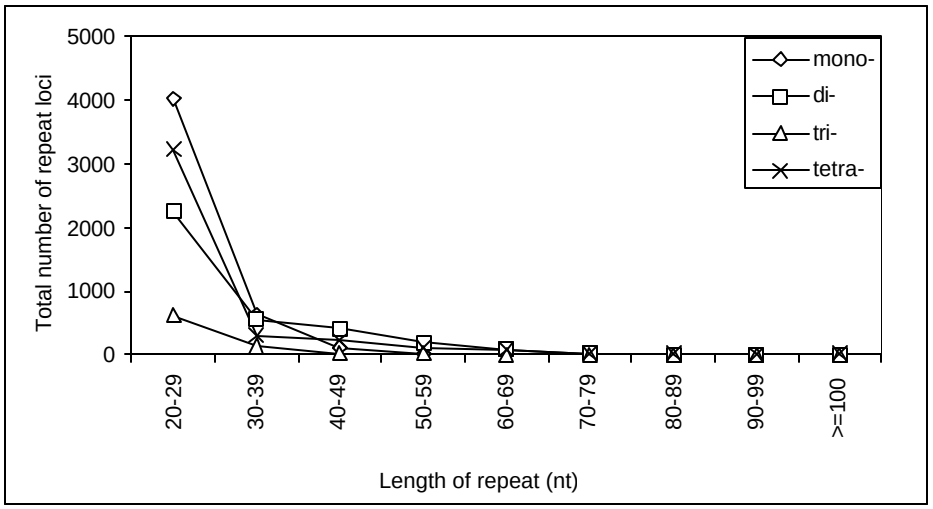


Figure 2.4 Frequency v/s length distribution of simple sequence repeats in human chromosome-21

The length distributions of all SSRs have indicated that the frequency of repeats decreases exponentially with the length of repeats (Figure 2.4 shows a representative graph of length distributions of repeats in human chromosome-21). This may be because longer repeats have higher mutation rates and hence are more unstable (Wierdl et al., 1997; Kruglyak et al., 1998). The paucity of longer microsatellites could be also due to their downward mutation bias and short persistence time (Harr and Schlotterer, 2000). Recent studies have shown that compared to expansion mutation events, contraction mutations occur more frequently with increase in allele size (Xu et al., 2000) and long alleles tend to mutate to shorter lengths, thus preventing their infinite growth (Ellegren, 2000a).

Among the repeats longer than ~40 nucleotides, the dinucleotide repeats were more frequent whereas, mononucleotide repeats seemed to be less common (Table 2.3). A large number of tetranucleotide repeats in human chromosomes and trinucleotide repeats in *Drosophila* were also longer than ~40 nucleotides. Slippage rates have been estimated to be the highest in dinucleotide repeats followed by tri- and tetranucleotide repeats (Kruglyak et al., 1998; Chakraborty et al., 1997; Schug et al., 1998). Probably, shorter repeating units allow more number of possible slippage events per unit length of DNA and hence, are likely to be more unstable. However, shorter lengths of mononucleotide repeats in all genome sequences and abundance of tetranucleotide repeats in human sequences suggest involvement of additional mechanisms.

My study shows that compared to human chromosome-21, chromosome-22 has significantly higher frequency of mono-, tri- and tetranucleotide repeats but less of dinucleotide repeats (*t* Test: t=5.60 for mono-, t=3.42 for di-, t=4.59 for tri- and t=3.94 for tetranucleotide repeats; $p < 0.01$ in all the cases). In *C. elegans*, among a total of 60 chromosome pairs / repeat type combinations, 15 combinations show significant difference in density of repeats (at $p < 0.05$). On the other hand, the densities of repeats in *Arabidopsis* chromosomes 2 and 4 are similar. In case of *Drosophila*, the sex chromosome (X) contains ~1.5 to 3 times more repeats per mbp of sequence as compared to autosomes (chromosome-2 and -3) (significance not calculated). Such differences for dinucleotide repeats in *Drosophila* sex chromosome and autosomes have been reported earlier (Pardue et al., 1987; Bachtrog et al., 1999). Thus, although the trends for different repeat classes are similar between

chromosomes within a genome, the density of repeats may vary between different chromosomes of the same species. This can be expected since different chromosomes in a genome can have different organization of genes, eu-chromatin and hetero-chromatin.

2.3.2 Relative frequencies of various di- and trinucleotide repeats:

All dinucleotide repeat combinations excluding homomeric dinucleotides can be grouped in four unique classes, namely $(AT)_n$, $(AG)_n$, $(AC)_n$ and $(GC)_n$. It is evident that, in human and *Drosophila* chromosomes, AC dinucleotide repeats are more frequent followed by AT and AG repeats (Figure 2.5). In contrast, *Arabidopsis* chromosomes contain more of AT repeats followed by AG repeats. However, in the yeast genome, AT repeats seem to be predominant compared to other dinucleotide repeats. Interestingly, GC dinucleotide repeats are extremely rare in all the genomes studied. Lower frequency of CpG dinucleotides in vertebrate genomes has been attributed to methylation of cytosine that in turn increases its chances of mutation to thymine by deamination (Schorderet and Gartler, 1992). However, CpG suppression by this mechanism can not explain the rarity of $(CG)_n$ dinucleotide repeats in yeast, *C. elegans* and *Drosophila* since they do not show cytosine methylation.

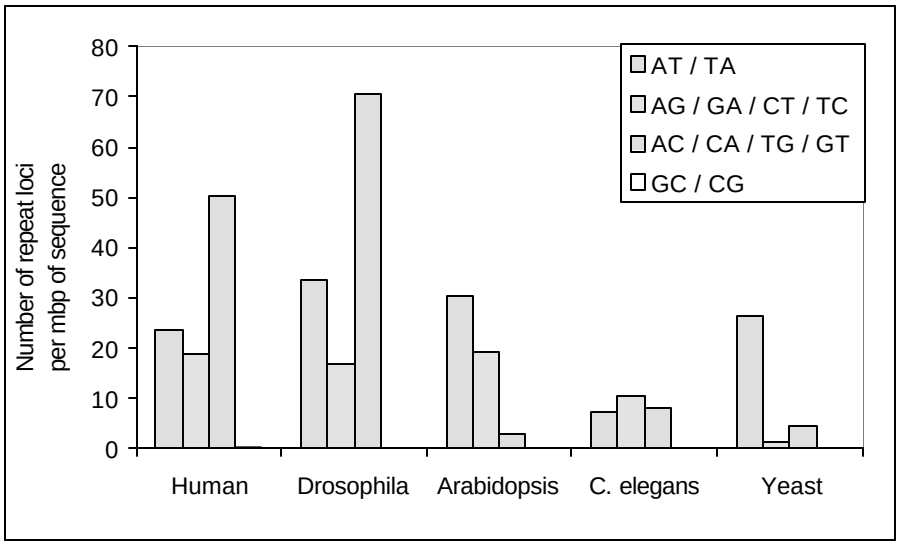


Figure 2.5 Frequency of different dinucleotide repeats per mbp of chromosome sequences in different genomes

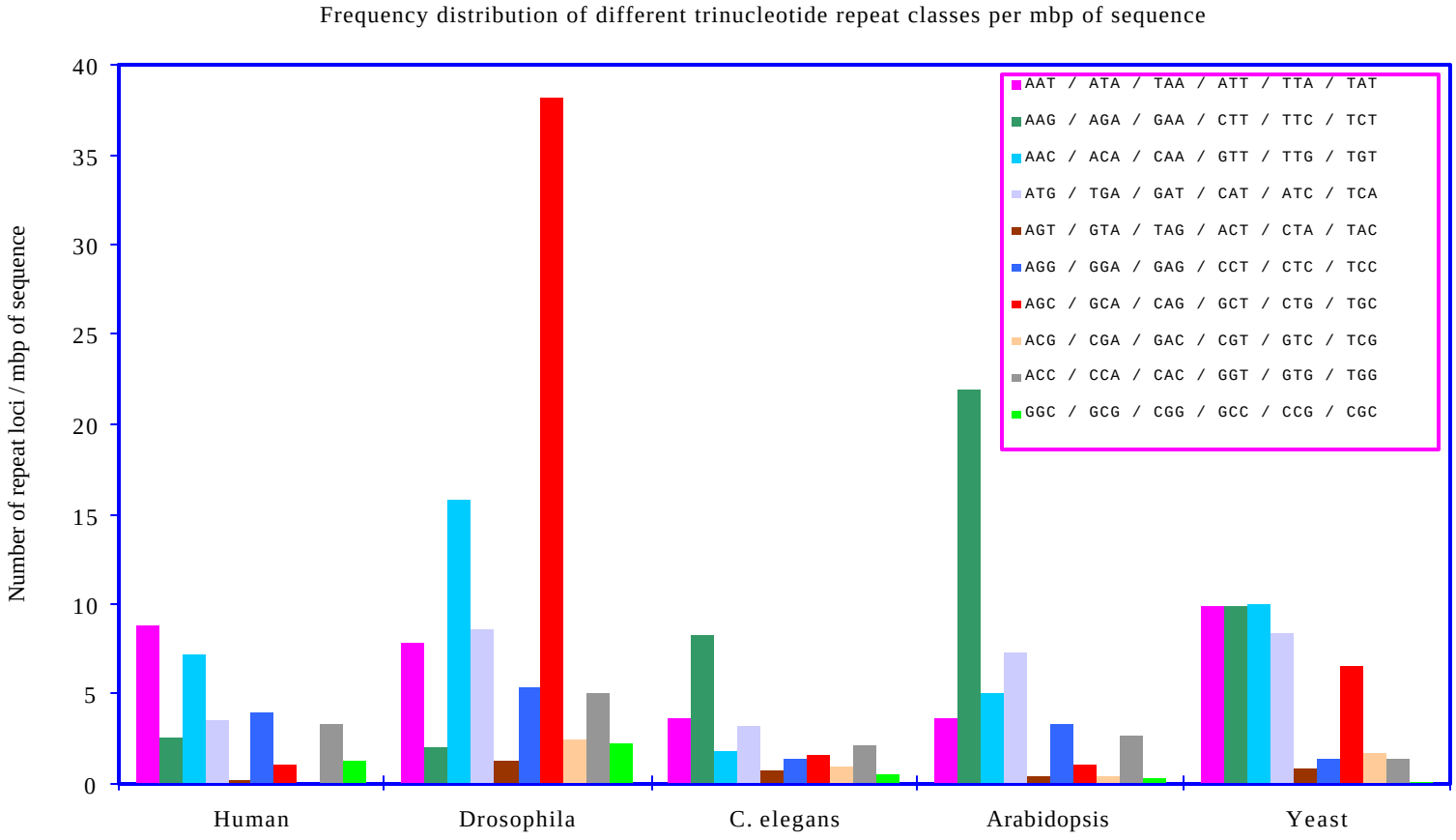


Figure 2.6 Frequency of different trinucleotide repeats per mbp of chromosome sequences in different genome

Among 10 unique trinucleotide repeat classes, human chromosomes 21 and 22 contain more of AAT and AAC repeats (Figure 2.6). Compared to other genomes, *Drosophila* chromosomes have the highest frequency of trinucleotide repeats and among them, AGC repeats are predominant followed by AAC repeats. The *Arabidopsis* and *C. elegans* chromosomes have comparatively higher frequency of AAG trinucleotide repeats. In contrast, yeast genome contains more of AAT, AAG, AAC, ATG and AGC repeats. It should be noted that frequencies of trinucleotide repeats in the chromosome sequences also include those occurring in the coding regions and could be partially limited by selection at protein level.

Short proto-microsatellites are probably generated by random mutations and then expand by DNA-slippage mediated events. Therefore, the base composition of a sequence that provides seeds for evolution of repeats is expected to influence microsatellite density (Bachtrog et al., 1999; Kruglyak et al., 2000). We tested this assumption first by XY-scatter plot representation of percent di- and trinucleotide composition of a sequence and frequency of corresponding repeats in individual chromosomes. Figure 2.7 is a representative graph of the relationship between percent nucleotide composition and frequency of di- and trinucleotide repeats in human chromosome-21. It was observed that differences in frequencies of various repeat classes were large and could not be attributed to differences in nucleotide composition of a sequence.

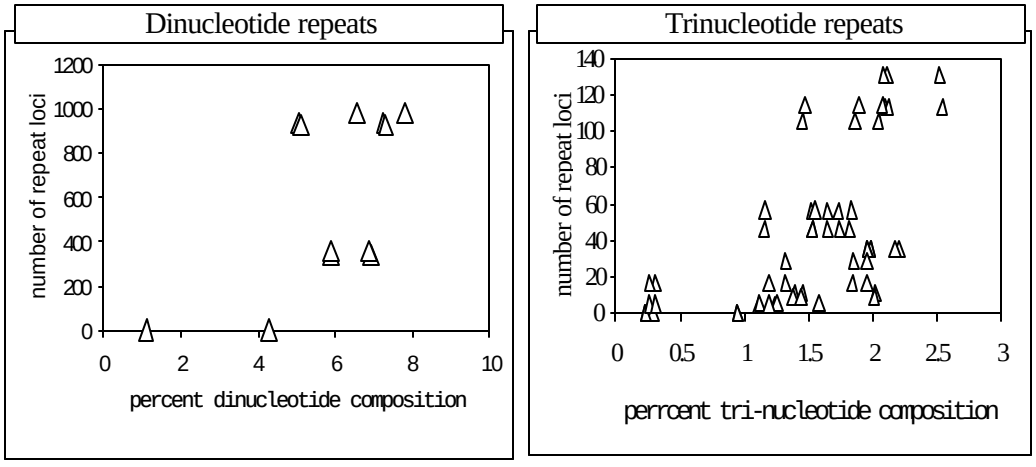


Figure 2.7 Scatter plot of data points showing relationship between percent nucleotide composition and frequency of repeats in human chromosome-21

DNA strand-slippage can occur during transient dissociation and re-annealing in the repeat region and this could be a deceptive event for DNA processing machinery leading to expansions or deletions in the repeat tracks (Figure 2.1). It has been suggested that if the nucleotides on the single strand are self-complementary, they can base pair to form loops or hairpins and stabilize strand slippage (Gacy et al., 1995; Moore et al., 1999). If these mechanisms favor repeat expansions / deletions, repeats with higher hairpin propensities like $(CTG)_n$, $(CCG)_n$ (Gacy et al., 1995; Mitas et al., 1995) or self-complementary repeats like $(AT)_n$, $(GC)_n$ are likely to be more abundant. However, relative frequencies of various di- and trinucleotide repeat classes within and between different genomes do not seem to support such an association. For example, AGC class of trinucleotide repeats (representing CAG / CTG repeats) are predominant in *Drosophila* whereas, in human, *Arabidopsis* and *C. elegans* genome sequences they are less frequent. In contrast, human chromosome-21 and -22 contain more of AAT and AAC trinucleotide repeats though their relative hairpin propensity is low (Gacy et al., 1995; Mitas et al., 1995). Similarly, AAG class of trinucleotide repeats that can adopt triple-helical structures (Pearson and Sinden, 1998) are comparatively more in *Arabidopsis*, *C. elegans* and yeast while they are less in human and *Drosophila* sequences. This suggests that, in addition to alternative DNA structures formed by repeat motifs, species specific cellular factors interacting with them are likely to play an important role in the genesis of repeats (Toth et al., 2000). It is likely that small sequence dependent differences in the efficiency of the enzymatic machinery of different genomes to detect and remove slippage mutations could result in vastly different mutation rates (Bachtrog et al, 2000) and characteristic microsatellite distributions.

2.3.3 Codon repetitions in complete genome coding DNA sequences:

Among all simple sequence repeats, slippage mediated expansions / deletions of only trinucleotide repeats or multiples thereof can be tolerated in coding regions since they do not disturb reading frame. I, therefore, analyzed the occurrences of codon (trinucleotide) repeats in coding DNA sequences of all the predicted peptides of *Drosophila*, *C. elegans* and yeast genomes (Tables 2.4 and 2.5). It is evident that codon repetitions are far more frequent in *Drosophila* compared to *C. elegans* that has in fact more predicted proteins than *Drosophila*. This is to be expected since the

frequency of microsatellites is very low in *C. elegans* (Figure 2.3). In *Drosophila* coding sequences, CAG codon (encoding glutamine) repetitions are predominant followed by AGC (serine), GAG (glutamic acid), GCA (alanine) and AAC (asparagine) repeats. On the other hand, in *C. elegans* coding sequences, GAT (aspartic acid), CCA (proline), CAA (glutamine), GAA (glutamic acid) and AAG (lysine) codon repeats are comparatively more frequent, though a very few of them are repeated for 14 or more times. In yeast ORFs (open reading frames), GAA (glutamic acid), CAA (glutamine), GAT (aspartic acid), AAT (asparagine) and CAG (glutamine) codon repeats are more in number. Such trends for triplet repeats in yeast ORFs have been also reported earlier and are thought to reflect functional selection acting on amino acid reiterations in the encoded proteins (Alba et al., 1999).

Table 2.4: Frequencies of trinucleotide repeat classes in genomic and coding sequences*

	Drosophila		*C. elegans*		Yeast	
	Genomic	Coding	Genomic	Coding	Genomic	Coding
Trinucleotide repeat class						
AAT / ATA / TAA / ATT / TTA / TAT	916	37	349	25	119	55
AAG / AGA / GAA / CTT / TTC / TCT	240	53	786	174	119	105
AAC / ACA / CAA / GTT / TTG / TGT	1850	259	167	103	120	105
ATG / TGA / GAT / CAT / ATC / TCA	1000	106	309	125	101	83
AGT / GTA / TAG / ACT / CTA / TAC	147	6	67	14	10	4
AGG / GGA / GAG / CCT / CTC / TCC	625	235	129	43	17	16
AGC / GCA / CAG / GCT / CTG / TGC	4470	1909	149	86	79	66
ACG / CGA / GAC / CGT / GTC / TCG	281	62	87	25	21	18
ACC / CCA / CAC / GGT / GTG / TGG	594	198	203	130	17	11
GGC / GCG / CGG / GCC / CCG / CGC	258	123	44	23	1	1
Total sequence length (mbp)	117.13	20.55	95.23	25.15	12.07	8.93
Total occurrences of repeats	10381	2988	2290	748	604	464

*$(AAA)_n$, $(TTT)_n$, $(GGG)_n$, and $(CCC)_n$ codon repeats in the coding sequences are not included here.

Table 2.5: Occurrences of codon repeats in complete genome Coding DNA Sequence (CDS) sets of *Drosophila*, *C. elegans* and yeast

Codons	Encoded amino acid residue	*Drosophila*		*C. elegans*		Yeast	
		Codon repeated for >= 7 times	Codon repeated for >= 14 times	Codon repeated for >= 7 times	Codon repeated for >= 14 times	Codon repeated for >= 7 times	Codon repeated for >= 14 times
GGA / GGG / GGC / GGT	Glycine	141	2	51	0	4	0
GCA / GCG / GCC / GCT	Alanine	274	14	49	0	13	0
GTA / GTG / GTC / GTT	Valine	4	0	7	0	3	0
CTA / CTG / CTC / CTT TTA / TTG	Leucine	12	0	8	1	3	1
ATA / ATC / ATT	Isoleucine	10	0	5	0	0	0
TGC / TGT	Cysteine	3	0	1	0	1	0
ATG	Methionine	4	0	0	0	0	0
TAC / TAT	Tyrosine	2	1	7	0	2	0
TTC / TTT	Phenylalanine	4	0	9	0	10	1
TGG	Tryptophan	0	0	0	0	0	0
CCA / CCG / CCC / CCT	Proline	54	0	103	0	7	0
TCA / TCG / TCC / TCT AGC / AGT	Serine	250	9	29	0	34	2
ACA / ACG / ACC / ACT	Threonine	119	3	32	0	4	0
AAC / AAT	Asparagine	175	10	25	1	79	16
CAA / CAG	Glutamine	1555	107	130	0	122	7
GAC / GAT	Aspartic acid	79	0	108	2	81	10
GAA / GAG	Glutamic acid	166	6	98	0	81	5
AAA / AAG	Lysine	47	0	78	0	22	0
CGA / CGG / CGC / CGT AGA / AGG	Arginine	2	0	9	2	4	1
CAC / CAT	Histidine	92	0	24	0	13	0
Total occurrences of repeats		2993	152	773	6	483	43
Total coding sequences analyzed		14080		19209		6283	
Total length of coding sequences (mbp)		20.55		25.15		8.93	

The correlation coefficient between frequencies of various trinucleotide repeat classes in coding sequences and in non-coding sequences (frequency in total genome sequences - frequency in total coding sequences) was found to be significant in *Drosophila* (r= 0.84, $p<0.01$) but insignificant in *C. elegans* (r=0.53) and yeast (r=0.37). It was also noted that within a trinucleotide repeat class, frequencies of different codon repeats vary considerably depending on the type of encoded amino acid. Perhaps, the interesting observation in my study is that, expansions of codons corresponding to small hydrophilic amino acids are tolerated more compared to hydrophobic amino acids and this is particularly evident for codons repeated for 14 or more times (Table 2.5). Therefore, while nucleotide composition might play an important role in genesis of repeats, in the coding sequences, their effect on structure and function of the encoded proteins would be a major selective force. For example, at DNA level, physical and chemical properties of $(AGC)_n$, $(GCA)_n$, $(CAG)_n$, $(CTG)_n$, $(TGC)_n$ and $(GCT)_n$ repeats are same and their frequencies can be expected to be comparable. However, in the *Drosophila* coding DNA sequence set, there are 204 occurrences of AGC (serine), 175 of GCA (alanine), 1480 of CAG (glutamine), 36 of GCT (alanine), 11 of CTG (leucine) and 3 of TGC (cysteine) codon repeats (codons reiterated for >= 7 times).

The trends observed for codon repeats in complete genome coding DNA sequences are consistent with my study of protein sequence database, where I observed that tandem single amino acid repeats of small hydrophilic amino acids are more frequent in proteins (Chapter-3, section 3.3.1). This might perhaps explain why majority of the repeat associated diseases are due to expansions of CAG repeats in specific genes. Since glutamine repeats are tolerated more in proteins, the initial small $(CAG)_n$ expansions in coding regions are likely to have enough survival value to remain in population. However, as their instability increases with increase in length, their effect on protein structure and function could be deleterious beyond a certain limit leading to malfunctioning of the protein (Perutz, 1999). On the other hand, initial small expansions of hydrophobic and basic amino acid residues could be lethal and hence would be eliminated from the population as soon as they appear. The availability of complete coding DNA sequence set of the Human Genome will enable us to test this hypothesis.

2.4 CONCLUSIONS:

Analysis of simple sequence repeats in genome sequences gives a snapshot of *in vivo* accumulated repeats. Overall, the trends observed for various repeat classes in genome sequences are in agreement with previous reports (e.g. Richard and Dujon 1996; Bachtrog et al., 1999; Kruglyak et al., 2000; Toth et al., 2000). However, with the availability of complete genome / chromosome sequences, we have begun to understand the extent to which repeats are generated in a genome. Differential distribution of various repeats observed in different genome sequences suggests that apart from nucleotide composition of repeats, the characteristic DNA replication/repair/recombination machinery might have an important role in the evolution of SSRs. In addition, their occurrence in coding regions seems to be limited by non-perturbation of reading frame and tolerance of expanding amino acid stretches in the encoded proteins. These observations have implications on our efforts to understand the instability of disease associated repeats.

Development of a web-resource on simple sequence repeats in eukaryotic genome sequences:

I have compiled the locations and sequences of all the microsatellite loci reported in this study in the form of a web-resource, which has been made available at the URL: http://www.ncl-india.org/ssr (Figure 2.8). This information could be useful for the selection of a wide range of microsatellite loci for studying their location and sequence dependent evolution. These loci could also be used as markers for the fine analysis of recombination events along individual chromosomes. Availability of microsatellite content of complete chromosome sequences should also facilitate comprehensive studies on direct role of microsatellites in genome organization, recombination, gene regulation, quantitative genetic variation and evolution of genes.

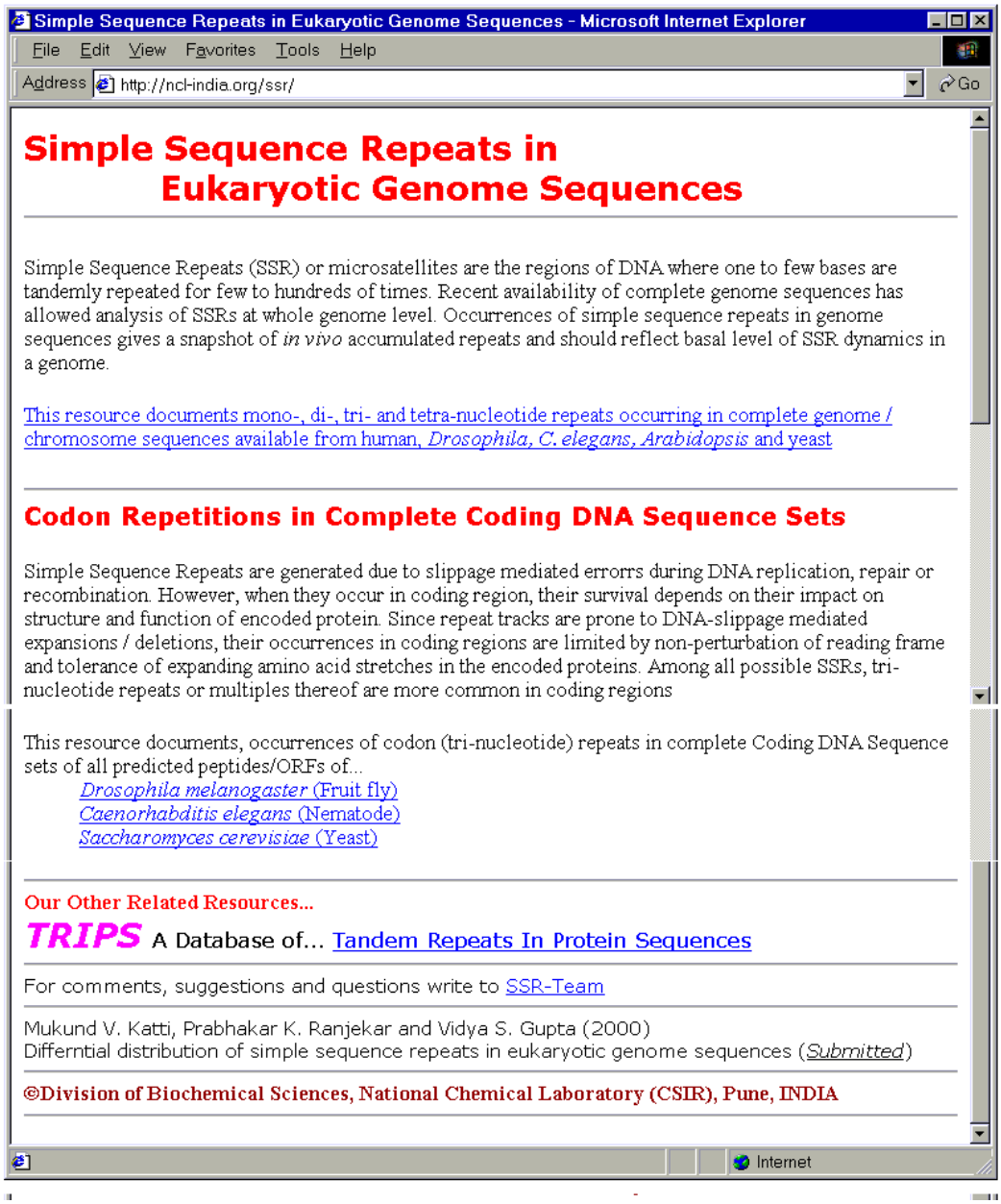


Figure 2.8: A web-resource on simple sequence repeats in eukaryotic genome sequences available at the URL: **http://www.ncl-india.org/ssr**

ABSTRACT:

All the protein sequences from SWISS-PROT database were analyzed for the occurrence of single amino acid repeats, tandem oligo-peptide repeats and periodically conserved amino acids. Single amino acid repeats of glutamine, serine, glutamic acid, glycine, and alanine seem to be tolerated to a considerable extent in many proteins. Tandem oligo-peptide repeats of different types with varying levels of conservation were detected in several proteins and found to be conspicuous, particularly in structural and cell surface proteins. It appears that repeated sequence patterns may be a mechanism that provides regular arrays of spatial and functional groups, useful for structural packing or for one to one interactions with target molecules. To facilitate further explorations, a database of *Tandem Repeats In Protein Sequences (TRIPS)* has been developed and made available at the URL: http://www.ncl-india.org/trips

3.1 INTRODUCTION:

In the previous chapter, I have described how simple sequence repeats of various types occur frequently in genome sequences. Although, SSRs originate due to errors during DNA replication/repair/recombination, when such events occur in the protein coding regions it might lead to appearance of repeated sequence patterns in proteins and can eventually dictate protein structure and function. In this chapter, I have included results of my studies on the occurrence of repeated sequence patterns in proteins and their implications on protein structure and function.

Redundancies in protein sequences have been noticed since the early days when DNA and protein sequencing techniques were established and protein sequence data started accumulating. Redundancy in a protein sequence can be in various forms, for example, as runs of identical amino acids or short tandem repeats or over-representation of certain amino acid combinations or partial gene duplications (Doolittle, 1989). Analysis of sequence databases have shown that single amino acid repeats are not rare in proteins and hydrophilic amino acids, particularly glutamine, account for a large proportion of single amino acid repeats (Green and Wang, 1994). Golding (1999) observed that in the yeast complete protein sequence set, the most common shared regions were runs of single amino acids or low complexity simple sequences rich in only one or a few amino acids. From the comparative analysis of yeast and several bacterial genomes, Pellegrini et al., (1999) have reported that eukaryotic proteins contain more internal repeats than those of prokaryotic or archeal organisms. They have also found that ~18% of yeast sequences and ~28% of the known human sequences contain detectable repeats indicating importance of internal duplications in protein evolution. Analysis of SWISS-PROT database has shown that duplicated sequence segments occur in ~14% of all proteins (Marcotte et al., 1999). The frequency distribution of repeats as a function of repeat length has revealed only weak length dependence suggesting recombination rather than duplex melting or DNA hairpin formation as the limiting mechanism underlying repeat formation.

Although internal repeats of various forms are known to occur in several proteins, occurrences of short tandem repeats in protein sequences have not been described systematically. Therefore, in order to present an overall picture of amino acid repeat patterns in protein sequences, I analyzed all the proteins from SWISS-PROT database (Bairoch and Apweiler, 1999) for the occurrence of single amino acid repeats, tandem oligo-peptide repeats and periodically conserved amino acids. I studied the observed repeat patterns in relation to their implications on protein structure and function. Moreover, I organized the results in the form of a database that has been made available through the Internet.

3.2 MATERIALS AND METHODS:

All the protein sequences from SWISS-PROT database (Release 38, of July 1999) were downloaded in FASTA format by ftp from the URL: ftp://expasy.ch and analyzed for various repeat patterns using computer programs written in "C" programming language. For detection of internal repeats in protein sequences, techniques like Fourier analysis (McLachlan, 1993) or modifications of dynamic programming algorithm (Heringa and Argos, 1993; Coward and Drablos, 1998; Pellegrini et al, 1999) have been applied. These algorithms simultaneously report repeat patterns of varying types occurring in a given sequence. Since I was interested to analyze complete protein database for tandem repeats of defined unit lengths and periodicity, I used a simple sliding window algorithm for detection of internal repeats in protein sequences. This allowed me to selectively search for repeats of defined lengths and classify them systematically. A brief description of the algorithm is outlined here.

3.2.1.1 Tandem single amino acid repeats:

Consider a protein sequence of length 's' as a string, $a_1a_2a_3a_4a_5$-----a_s, where, 'a_i' is amino acid residue at position 'i' in the sequence space. To detect tandem single amino acid repeat of a minimum length 'n', starting at position 'i', we compare amino acid 'a_i' with each of the subsequent residues 'a_{i+1}', 'a_{i+2}', 'a_{i+3}', ... 'a_{i+n-1}'. If all of them match, a repeat is detected and further extended as long as '$a_{i+n-1+j}$' (where j = 1, 2, ...) matches with 'a_i'. All the protein sequences were searched for tandem single amino

acid repeats of length >=5, >=10 or >= 15 without allowing mismatch and for repeats of length >=20 by allowing a maximum of 1 mismatch in 10 residues.

3.2.1.2 Tandem oligo-peptide repeats:

Consider an oligo-peptide of size 'k' as a window '$a_i a_{i+1} \ldots a_{i+k-1}$', in a protein sequence (e.g. k=2 for di-peptide repeats, k=5 for penta-peptide repeats). This window is compared with subsequent windows starting at positions, 'a_{i+k}', 'a_{i+2k}', 'a_{i+3k}',..... '$a_{i+(n-1)k}$'. An oligo-peptide repeat is identified and further extended if a minimum 'n' number of windows match with each other allowing a certain degree of mismatch (Table 3.1). While scanning for long oligo-peptide repeats, oligo-peptide units representing perfect repeats of shorter length were ignored. A representative flow-chart diagram describing the algorithm used for detection of penta-peptide repeats in protein sequences is illustrated in Figure 3.1.

3.2.1.3 Periodic conservation of single amino acids:

The protein sequences were scanned for periodic conservation of single amino acids essentially using the same algorithm as used for detection of single amino acid repeats, except that 'a_i' is compared with 'a_{i+p}', 'a_{i+2p}', 'a_{i+3p}', 'a_{i+4p}'..... '$a_{i+(n-1)p}$' where 'p' is period of 2 to 10.

3.2.2 Development of the database:

One of the objectives of my work was to develop a comprehensive database of short tandem repeats in protein sequences. I designed the computer programs in such a way that the outputs were automatically written in HTML (Hyper-Text Markup Language) format. Using HTML, simple text information can be enriched by inserting tags that allow display of information in attractive fashion using suitable fonts, colors, tables and images, when viewed through an appropriate browser. More importantly, hyperlinks can be provided to other documents on any computer connected to the net and thereby, users can easily retrieve additional information and explore.

Table 3.1: Minimum repeating units and mismatch parameters used for detection of various oligo-peptide repeats.

Oligo-peptide unit length	Minimum repeating units	Maximum mismatch allowed	Number of proteins containing the repeats
2	7	10%	161
3	5	10%	109
4	4	10%	117
5	4	10%	76
6	4	10%	74
7	4	10%	58
8	4	10%	108
9	3	10%	58
10	3	10%	114
11	3	15%	41
12	3	15%	119
13	3	15%	32
14	3	15%	64
15	3	15%	51
16	3	20%	120
17	3	20%	32
18	3	20%	85
19	3	20%	37
20	3	20%	113

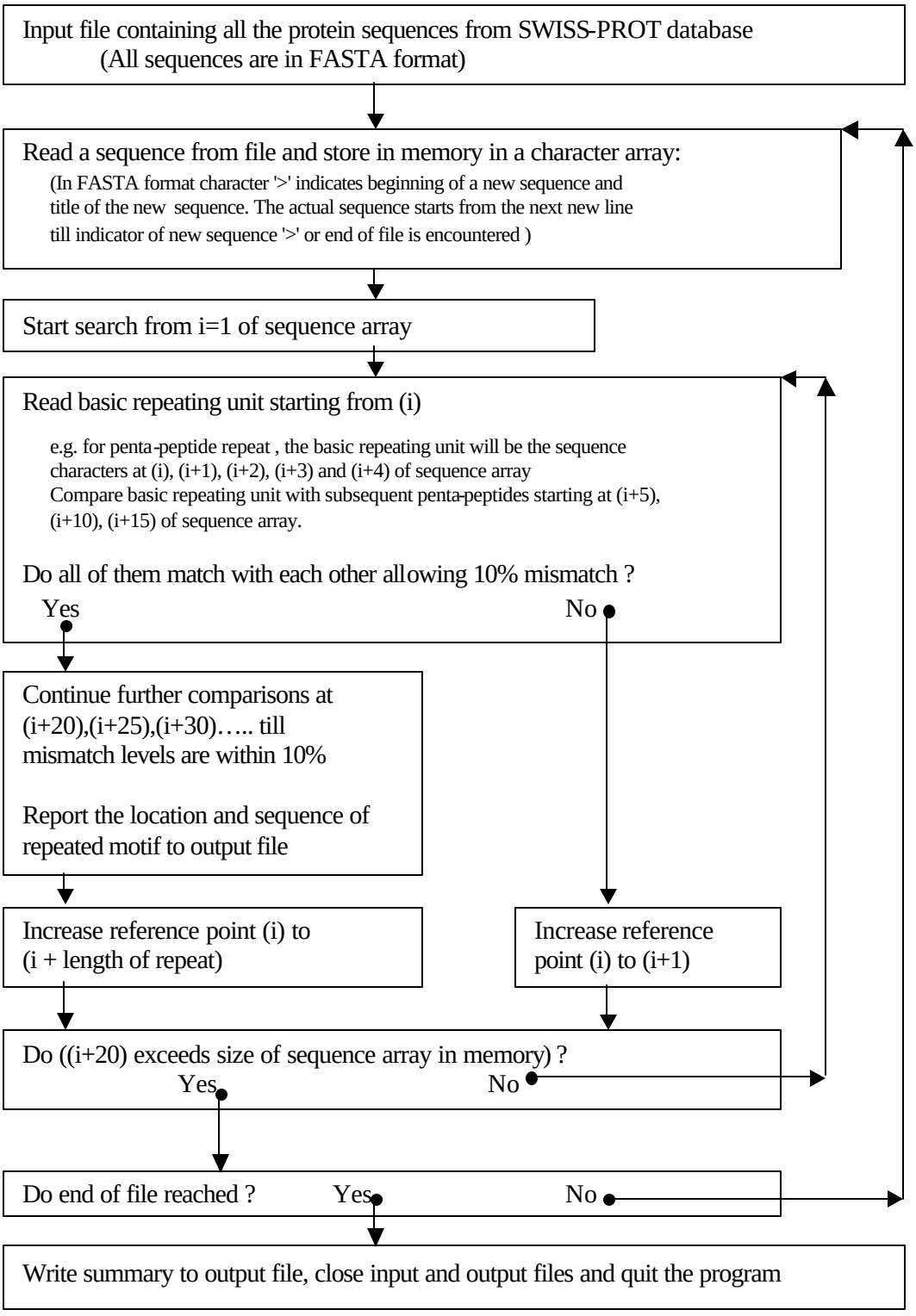


Figure 3.1: A representative flowchart showing the algorithm used for detection of penta-peptide repeats in protein sequences.

I compiled and organized the results of my analysis in the form of a database of *Tandem Repeats In Protein Sequences* (TRIPS) that has been made available at the URL: http://www.ncl-india.org/trips. The *TRIPS* database is organized in 3 major sections describing single amino acid repeats, tandem oligo-peptide repeats and periodically conserved amino acids (Figure 3.2). In the individual files, each protein entry describes SWISS-PROT accession number, protein name, sequence length, type and length of repeat, position of repeat in the sequence and actual repeat pattern. Hyperlink is provided to the original entry in the SWISS-PROT database and thereby further information about the protein sequence, domain structure, function and relevant literature can be easily searched. For tandem oligo-peptide repeats, I have used different color schemes for different types of amino acids (hydrophobic, hydrophobic-aromatic, polar uncharged, acidic, basic and unique) which help in effective visual display of repeat patterns.

3.3 RESULTS AND DISCUSSION:

For this analysis, I have selected SWISS-PROT database since it has minimal redundancy and protein sequence entries have rich annotations and extensive links to other databases (Bairoch and Apweiler, 1999). All the 80,000 proteins from the database were analyzed for tandem repeats using a simple sliding window technique with empirically determined mismatch levels and repeat cut-off units. Although such a method may not detect distant repeats, repeats falling beyond cut-off scores or those containing insertions or deletions, it was possible to present an overall picture of repeat patterns observed in the database.

3.3.1 Tandem single amino acid repeats:

Table 3.2 summarizes total number of proteins containing single amino acid repeats of various types. The complete lists of proteins containing tandem single amino acid repeats are available through the *TRIPS* database whereas, representative proteins are described in Table 3.3.

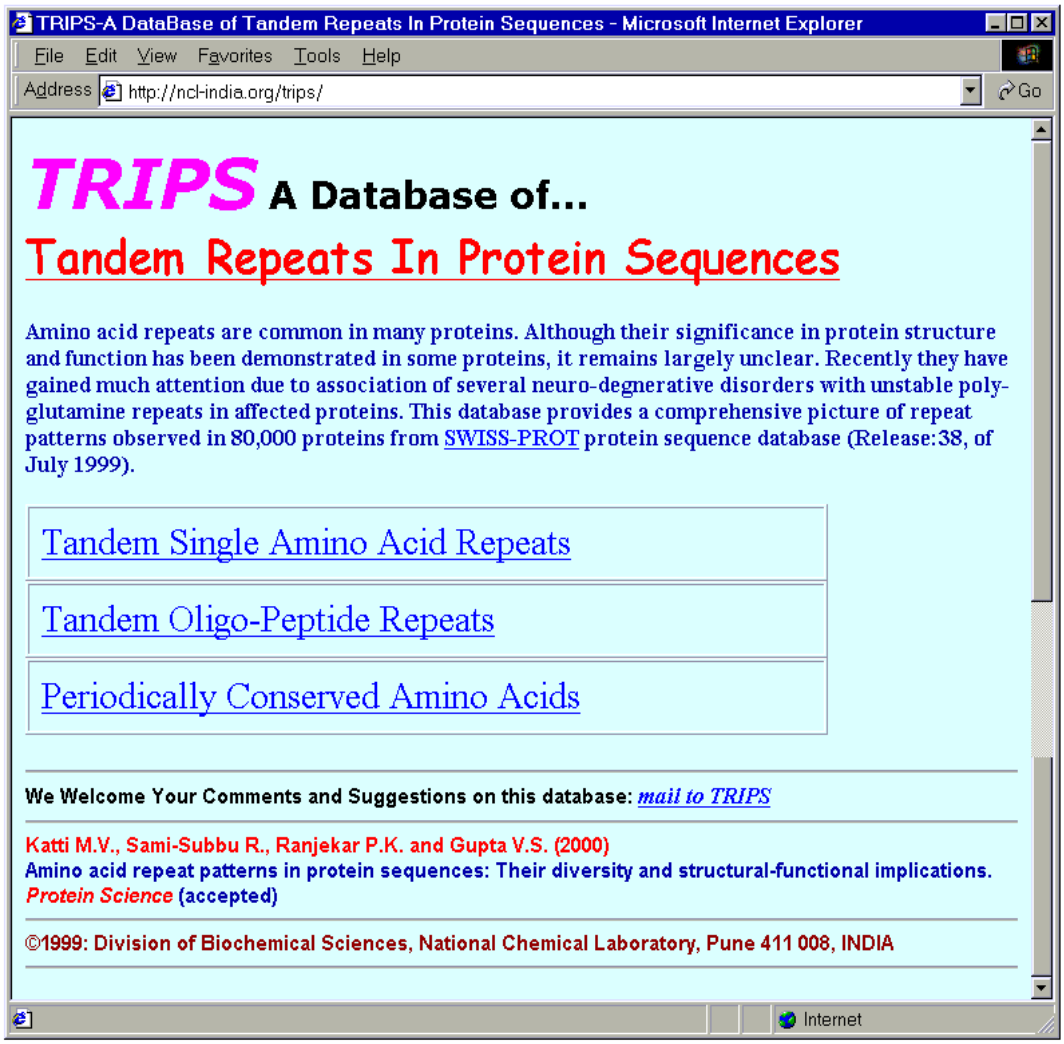


Figure 3.2: Homepage of the database of *Tandem Repeats In protein Sequences* (TRIPS) available at the URL: **http://www.ncl-india.org/trips**

Table 3.2: Number of proteins in the SWISS-PROT database containing tandem single amino acid repeats

Amino acid	Amino acid frequency in the database (%)	Number of proteins containing repeats			
		Repeat length >= 5	Repeat length >= 10	Repeat length >= 15	Repeat length >= 20 allowing 10% mismatch
Small					
Glycine	6.84	679	76	24	29
Alanine	7.58	1078	87	16	8
Hydrophobic					
Valine	6.58	61	0	0	0
Leucine	9.44	1019	10	2	0
Isoleucine	5.81	37	0	0	0
Cysteine	1.66	8	1	0	0
Methionine	2.38	11	0	0	0
Tyrosine	3.19	11	1	0	0
Phenylalanine	4.10	22	2	0	0
Tryptophan	1.24	0	0	0	0
Proline	4.92	630	39	11	11
Hydrophilic, uncharged					
Serine	7.13	977	70	25	29
Threonine	5.68	270	25	3	6
Asparagine	4.44	257	50	23	29
Glutamine	3.97	622	173	73	74
Hydrophilic, acidic					
Aspartic acid	5.28	341	26	11	10
Glutamic acid	6.37	804	74	20	24
Hydrophilic, basic					
Lysine	5.95	289	2	0	0
Arginine	5.16	326	5	0	0
Histidine	2.25	171	40	0	0

Table 3.3: Representative proteins containing long single amino acid repeats

Accession number	Amino acid	Repeat length[a]	Organism	Protein name
P18480	Glutamine	51/56	*S. cerevisiae*	Transcription Regulatory Protein SNF-5
P54683	Glutamine	42/46	*D. discoideum*	Prestalk-Specific Protein TAGB
P20226	Glutamine	40/41	*H. sapiens*	Transcription Initiation Factor TFIID
O61735	Glutamine	40/43	*D. melanogaster*	Circadian Locomoter Output Cycles Kaput Protein (DCLOCK)
P54637	Asparagine	53/58	*D. discoideum*	Protein-Tyrosine Phosphatase 3
P54674	Asparagine	42/42 + 36/39	*D. discoideum*	Phosphatidylinositol 3-Kinase 2
P54683	Asparagine	36/38	*D. discoideum*	Prestalk-Specific Protein TAGB Precursor
P32583	Serine	61/68	*S. cerevisiae*	Suppressor Protein SRP-40
P18709	Serine	35/38 + 28/28	*X. laevis*	Vitellogenin A2
P42568	Serine	44/46	*H. sapiens*	AF-9 protein
P31231	Aspartic acid	44/44	*R. esculenta*	Calsequestrin, Skeletal Muscle Isoform
P13816	Glutamic acid	36/39	*P. falciparum*	Glutamic Acid-Rich Protein
P19351	Glutamic acid	34/36	*D. melanogaster*	Troponin T, Skeletal Muscle
P21997	Proline	43/48	*V. carteri*	Sulfated Surface Glycoprotein 185
P12978	Proline	40/42	*Epstein-barr virus*	EBNA-2 Nuclear Protein

[a] repeat length is represented as 'm/n' indicating 'm' number of repeated single amino acids in a stretch of 'n' residues

Table 3.4: Single amino acid repeats in solved structures

PDB ID	Single amino acid repeats[a] and corresponding secondary structures[b]	Sequence length
1L64	35 KSPSL AAAAAAAAAA IGRN SSS*H HHHHHHHHHH HTS*	164
1C9R:B	235 HPDKW AAAAAAAAAAAAA TVNDI **SS* **********TT* *HHHH	430
168L:A	123 QKRWD AAAAALAAAA WYNQT TT*TT HHHHHHHTTH HHHHT	164
1QGN:A	2 AKAVD AAAAAAAAIA PVDTT ***** ********** *****	445
1WFA:A	1 DTASD AAAAAA LT AANAKAAA ELT AANAAAAAAA TARX -(C-Terminal) **HHH HHHHHH HH HHHHHHHH HHH HHHHHHHHHH HH**	38
1FPV	23 SGNGS GGGGGGGSGG VGISTG ***** ********** TTS***	584
4DPV:Z	23 SGNGS GGGGGGGSGG VGIST ***** ****EETTEE TT***	584
1SPF	11 LKRLL VVVVVVVLVVVVIV GALLM HHHHH HHHHHHHHHHHHHH HHHHH	35
1AYZ:A	149 WEDDM DDDDDDDDDDDD EAD -(C-Terminal) HHHHT ************ ***	169
1A8Y	350 EINTE DDDDEDDDDDDD -(C-Terminal) ***** ************	367
1FT1:A	17 GQPEQ PPPPPPPPP AQQPQ ***** ********* *****	377
1BN5	93 DGATG KKKKKKKKK RGPKV ***** ********* *****	148

[a] repeat length >=10 with a maximum of 1 mismatch in 10 residues

[b] Secondary structure information was obtained from Protein Data-Bank (PDB) web-site http://www.rcsb.org/pdb ; PDB-secondary structure element codes are: H-alpha helix (4-helix); E-extended strand, participates in beta ladder; G-310 helix (3-helix); T- hydrogen bonded turn; S-bend; * no regular secondary structure

Among the proteins containing 10 or more repetitions of identical amino acids, glutamine, alanine, glycine, glutamic acid, and serine repeats were much more frequent than other amino acids. Interestingly, no protein in the database contained tryptophan consecutively repeated for 5 or more times. Similarly, a very few repeats of cysteine, methionine, tyrosine, phenylalanine, isoleucine, and valine were detected, suggesting that long tandem repeats of highly hydrophobic amino acids are probably not favored in proteins. Such trends have also been reported by Green and Wang (1994).

Of the proteins containing long tandem single amino acid repeats (n>=10), more than $1/3^{rd}$ were transcription regulatory proteins with particularly more frequent poly-glutamine or poly-alanine repeats. Poly-glutamine rich regions in transcription factors are possibly involved in modulation of transcription activation (Gerber et al., 1994). Synthetic poly-glutamine peptides have been shown to form beta-sheets and might function as polar-zippers in protein-protein interaction (Perutz et al., 1994; Figure 3.8).

Poly-glutamine repeats encoded by CAG codons are found to be unstable, since the corresponding $(CAG)_n$:$(CTG)_n$ repeats in the coding sequences can readily adopt unusual DNA structures leading to errors during replication, repair or recombination (Pearson and Sinden, 1998). Such dynamic poly-glutamine repeat expansions in affected proteins have been shown to cause several neuro-degenerative disorders (Table 2.1). The common mechanism in these diseases seems to be misfolding of affected proteins with expanded glutamine stretches, formation of insoluble aggregates or intra-nuclear inclusions and eventual neuronal death (Paulson, 1999; Perutz, 1999). In addition to 'poly-glutamine diseases' there is at least one more example of a disease caused by expansion of single amino acid repeats. Muragaki et al., (1996) have reported that expansion of poly-alanine repeat in human homeo-box protein HOX-D13 leads to synpolydactyly, which is characterized by abnormality of hands and feet. These observations point out the severity of unstable single amino acid repeats and their significance in bio-medical research.

Analysis of SWISS-PROT entries of proteins containing tandem single amino acid repeats (n>=10) indicates that single amino acid repeats have not been assigned clearly to any functional domains. In several cases, single amino acid repeats show length variations in the same protein across species. For example, the TATA box binding protein (TF IID) contains a poly-glutamine region in its N-terminal domain that consists of 38 consecutive glutamine residues in Human, whereas 14 in hamster, 13 in mouse, 6 in chicken, 6 in viper and 4 in *Xenopus* (Figure 3.3). Thus, single amino acid stretches may not serve any function and may be only a mechanism for increasing the size of the protein (Green and Wang, 1994). The protein structures are intrinsically stable at domain level and show considerable flexibility in terms of sequence or length of short linker groups (Heringa and Taylor, 1997). Therefore, most probably these repeats occur in the linker regions and their probable function may be to simply serve as spacers between the domains (Golding, 1999; Huntley and Golding, 2000). It can be speculated that expansions of amino acid repeats, particularly hydrophilic amino acids, could be tolerated to a considerable extent if they occur in the linker regions and if they can be easily solvated on surface of the protein.

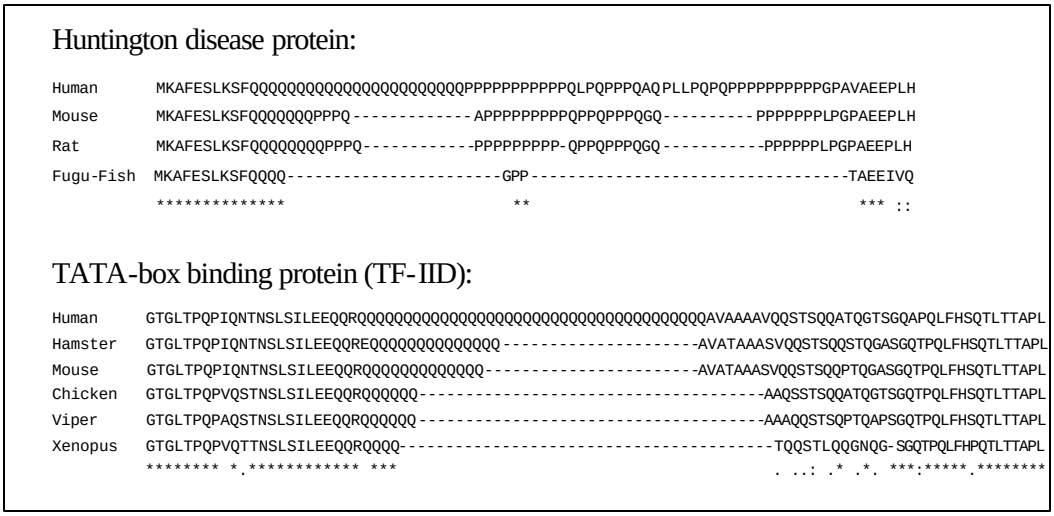


Figure 3.3: Sequence alignment of poly-glutamine region of Huntington Disease protein and TATA-box binding protein from different organisms

Since protein crystal structure studies can give direct insights into implications of amino acid repetitions on protein secondary and tertiary structure, I scanned the protein sequences from the PDB database for single amino acid repeats of length >= 10 allowing a mismatch of 1 in 10 residues. Table 3.4 describes the detected single amino acid stretches along with their reported secondary structures. It is evident that poly-alanine stretches can form regular alpha helix as well as combinations of alpha helix, bends, turns and also non-regular structures. As observed in pulmonary surfactant associated polypeptide-C (PDB-ID 1SPF), all hydrophobic poly-valine track formed a single alpha helix. In Calsequestrin (PDB-ID 1A8Y) and Ubiquitin-conjugating enzyme Rad6 (PDB-ID 1AYZ), the poly-aspartate repeats did not form any regular secondary structures possibly because they happened to be in the C-terminal domains. It appears that single amino acid repeats may adopt regular as well as non-regular structures and this could be largely influenced by their hydrophobicity / hydrophilicity and their context in the parent protein.

3.3.2 Tandem oligo-peptide repeats:

Protein sequences from the SWISS-PROT database were further scanned for tandem oligo-peptide repeats of length 2 to 20. Table 3.1 summarizes total number of proteins containing oligo-peptide repeats of various types whereas representative examples of proteins are listed in Table 3.5. Figures 3.4 and 3.5 show a few interesting oligo-peptide repeat patterns observed in protein sequences.

Among the proteins containing long oligo-peptide repeats, the antigenic proteins from malarial parasite, *Plasmodium*, showed a wide range and high sequence polymorphism (Figure 3.5). They include circumsporozoite protein, sporozoite surface protein, merozoite surface antigen, ring-infected erythrocyte surface antigen, duffy receptor (erythrocyte binding protein), malarial antigen P101 and s-antigen protein. Except the s-antigen protein, all others are cell surface proteins and are involved in interaction with the host cells (Holder, 1994). In another protozoan parasite, *Trypanosoma*, the shed antigenic proteins (trans-sialidase) contain extensive 12 residue repeats that act as immunomodulator and stabilize sialidase activity of the protein (Buscaglia et al., 1999).

Table 3.5: Representative proteins containing long tandem oligo-peptide repeats

Oligo length	Accession number	Oligo-peptide repeating unit	Number of repeats[a]	Organism	Protein name
2	P19275	PT	45+27+10	*TTV 1*	Viral Protein TPX
	P10220	PQ	35	*HSV*	Large Tegument Protein
	P14922	QA	32	*S. cerevisiae*	Glucose Repression Mediator Protein
	P07663	GT	29	*D. melanogaster*	Period Circadian Protein
3	Q01443	PNN	26	*P. berghei yoelii*	Sporozoite Surface Protein 2
	P54705	DSR	15	*D. discoideum*	Putative Chromatin Binding Protein-SNWA
	P07916	GVP	13	*G. gallus*	Elastin
4	P14593	AAGN	66	*P. brasilianum*	Circumsporozoite Protein
	P08307	PNAN	43	*P. falciparum*	Circumsporozoite Protein
	P22699	TETP	21	*D. discoideum*	Endoglucanase
5	P02840	PTTTK	23	*D. melanogaster*	Salivary Glue Protein SGS-3
	P13730	XTKRA	16	*D. erecta*	Salivary Glue Protein SGS-3
	P04985	PGVGV	11	*B. taurus*	Elastins A/B/C
6	P19246	EAKSPX	9+30	*M. musculus*	Neurofilament Triplet H Protein
	P08675	DGARAE	19	*P. cynomolgi*	Circumsporozoite Protein
	P05790	SGAGAG	16	*B. mori*	Fibroin Heavy Chain
7	P24928	SPSYSPT	47	*H. sapiens*	RNA Polymerase II Largest Subunit
	Q00725	TEPPXCX	12+8	*D. melanogaster*	Salivary Glue Protein SGS-4
	P32323	TSXSSTS	17	*S. cerevisiae*	A-Agglutinin Attachment Subunit
8	P13821	GPNSDGDK	66	*P. falciparum*	S-Antigen Protein
	P24587	TVGQAEEA	21	*R. norvegicus*	A-Kinase Anchor Protein 150
	P10419	GREXQGRF	18	*A. elegantissima*	Antho-Rfamide Neuropeptide Precursor
	Q62267	PEPCHPKA	12	*M. musculus*	Small Proline-Rich Protein (Cornifin-B)
9	Q03110	GDRADGQPA	21	*P. simium*	Circumsporozoite Protein
	P42565	XXDPFLRFG	13	*L. stagnalis*	FmRfamide-Related Neuropeptide Precursor
	P10667	TTPETTTVP	12	*X. laevis*	Integumentary Mucin A.1
10	Q40375	PPVYKPPVEK	33	*M. truncatula*	Repetitive Proline-Rich Cell Wall Protein
	P07476	XEQQEGQLEL	11	*H. sapiens*	Involucrin
	Q14242	XEAQTTXXAA	10	*H. sapiens*	P-Selectin Glycoprotein Ligand 1

Table 3.5: continued…

Table 3.5: continued…

11	P09593	GGPGSEGPKGT	19	*P. falciparum*	S-Antigen Protein
	P19835	PVPPTGDSXXX	16	*H. sapiens*	Bile-Salt-Activated Lipase
	P08674	DGAAAAGGGGN	14	*P. cynomolgi*	Circumsporozoite Protein
12	P23253	DSSAHXTPSTPX	44	*T. cruzi*	Sialidase
	P97347	QSXHXGQKGRXD	14+5+4	*M. musculus*	Repetin
	P13813	TEETQKTVEPEQ	13	*P. knowlesi*	110 Kd Antigen
	Q28824	TPKPLXXXKPAE	13	*O. cuniculus*	Myosin Light Chain Kinase, Smooth Muscle
13	P12027	ATSEAATGPSGDD	33	*O. mykiss*	Apopolysialo-glycoprotein
	P10547	AEVETSKAPVENT	15	*S. simulans*	Lysostaphin
14	P41809	SXPXAXSSTYTSSP	24	*S. cerevisiae*	HM1 Killer Toxin-Resistant Protein
	P05143	QGPPPPGGPQPRPP	13	*M. musculus*	Proline-Rich Protein MP-3
	P12036	KSPEKAKSPXKXEA	9	*H. sapiens*	Neurofilament Triplet H Protein
15	P08021	VDKRFMRFGKSVDGD	10	*A. californica*	FMRFAMIDE Neuropeptide Precursor
	Q60557	ETXTTVGNQSVTPGG	7	*M. auratus*	Oviduct-Specific Glycoprotein
16	Q08696	AKKXKEXXEXKXCXXX	23+10+16	*D. hydei*	Axoneme-Associated Protein MST101(2)
	P09815	LXAGYGSTXTAXXXSX	45	*P. fluorescens*	Ice Nucleation Protein
	Q99102	XSSXSXGHATXLPVTD	13+23	*H. sapiens*	Tracheobronchial Mucin-4
	P24587	QAEEATVGXXXXATVX	12	*R. norvegicus*	A-Kinase Anchor Protein 150
	P21917	APXLPXXPCGPDCAPP	7	*H. sapiens*	D(4) Dopamine Receptor
17	P24856	AATAATXATXATXAXXF	46	*N. coriiceps*	Antifreeze Glycopeptide Polyprotein Afgp7/Afgp8
	P32334	SQVSDTXVXXTXSXSSV	7	*S. cerevisiae*	MSB-2 protein
18	P02674	TGSXXGGSWXTGGRTEPN	22	*P. marinus*	Fibrinogen Alpha-1 Chain
	Q03180	KSTAAXVSQIXDGQVQAA	8	*S. cerevisiae*	Covalently-Linked Cell Wall Protein 8 (PIR-3)
19	P16112	LETXAPGVEXISGLPSGEV	22	*H. sapiens*	Aggrecan Core Protein Precursor
	Q03178	XAXXSQIGDGQIQATTXTX	8*	*S. cerevisiae*	Covalently-Linked Cell Wall Protein 6 (PIR-1)
20	P15941	PPAHGVTSAPDTRPAPGSTA	43	*H. sapiens*	Polymorphic Epithelial Mucin-1
	P07898	PEIXXEXSTXXEXXGEXSAX	18	*G. gallus*	Aggrecan Core Protein
	P26907	KGGEXTSXNHDKEFYQEIGX	5	*B. subtilis*	Glucose Starvation-Inducible Protein B
	P06680	RPPKPGNQXGPPQQEGQQQN	5	*M. auratus*	Acidic Proline-Rich Protein

[a] allowed mismatch levels are as described in Table 3.1

* contains a frame-shift within the repeat

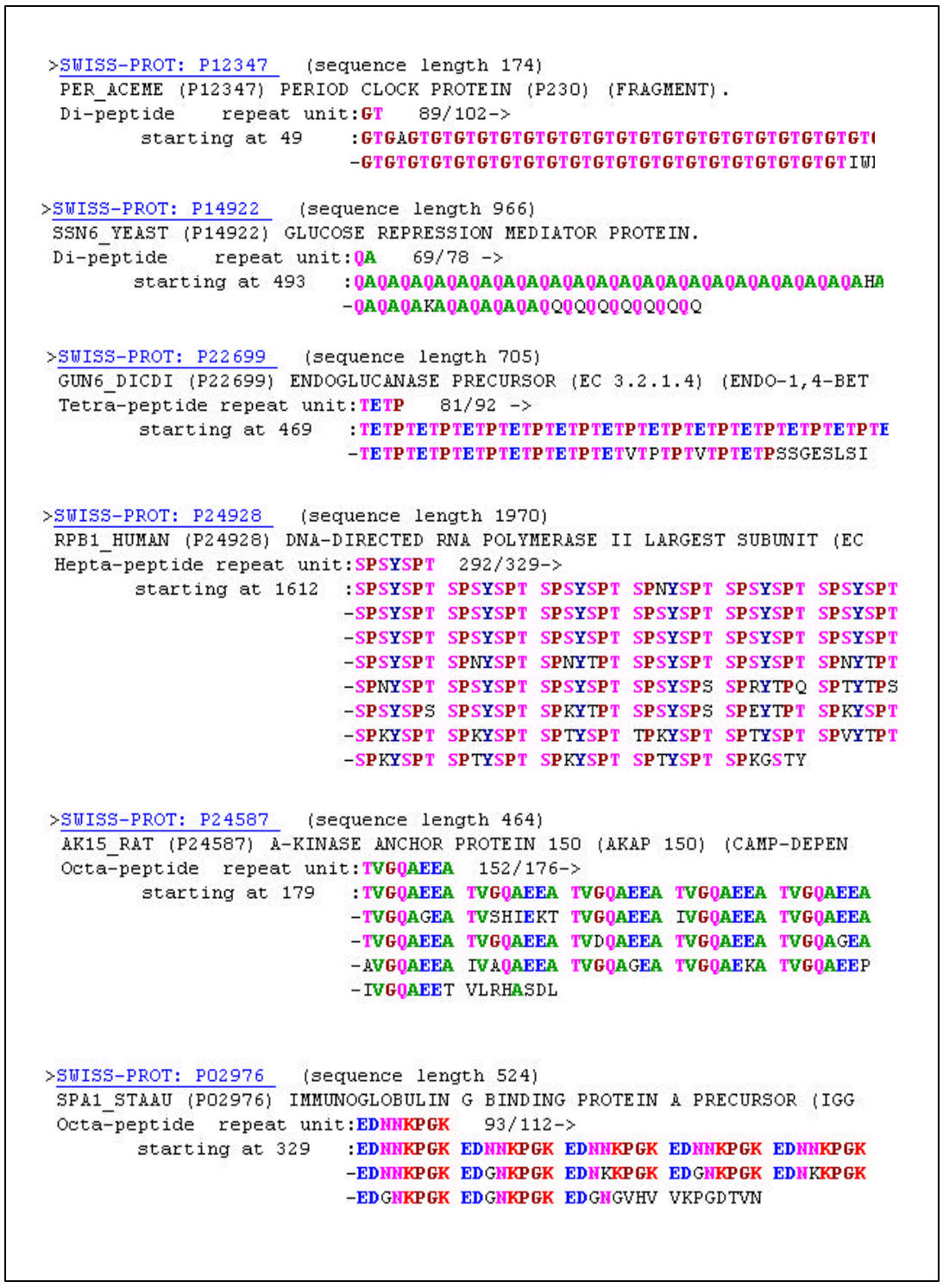


Figure 3.4a: A few examples of protein sequences containing tandem oligo-peptide repeats

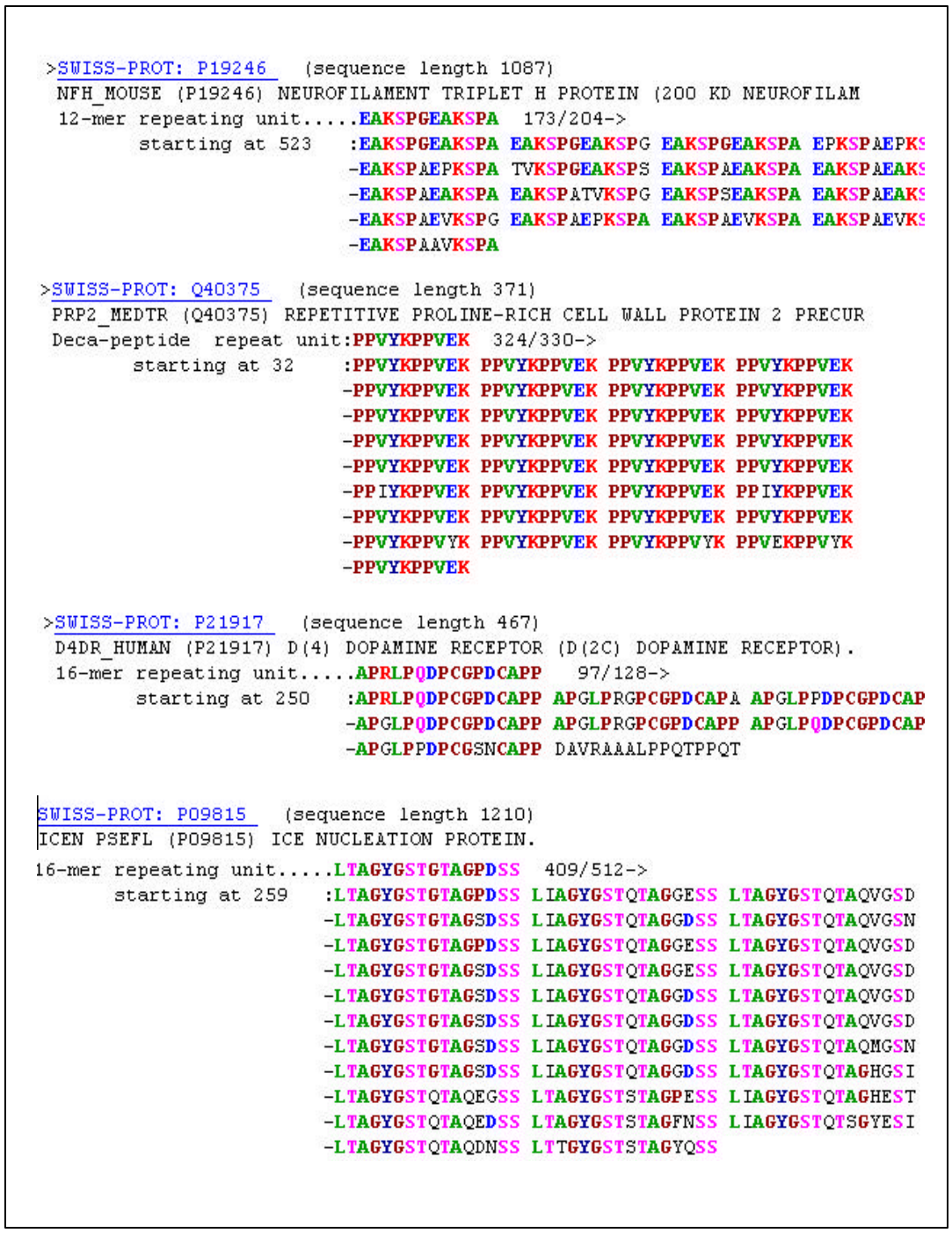


Figure 3.4b: A few examples of protein sequences containing tandem oligo-peptide repeats

```
>SWISS-PROT: P08676   (sequence length 419)
 CSP_PLACM (P08676) CIRCUMSPOROZOITE PROTEIN PRECURSOR (CS).
 Tetra-peptide repeat unit:GNAG  209/236->
        starting at 102   :GNAGGNAGGNAGGNAGGNAGGNADGNAGGNAGGNAGGNAGGNAGGN
                          -GNAGGNAGGNAGGNADGNAGGNAGGNAGGNADGNAGGNAGGNAGGN
                          -GNAGGNAGGNAGGNADGNAGGNAGGNAGGNADGNAGGNAGGNAGGN
                          -GNAGGTAGGNADGNAGGNAGGNAGGNAGGNAGGNAGGNAGGNAGGN
                          -GNAGGNAGGNAGGNAGANAGNKKAGDAGAGQGQNNEAANMPNVK

>SWISS-PROT: P13821   (sequence length 640)
 SANT_PLAFW (P13821) S-ANTIGEN PROTEIN PRECURSOR.
 Octa-peptide  repeat unit:GPNSDGDK  528/544->
        starting at 90    :GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK
                          -GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK GPNSDGDK

>SWISS-PROT: P08677   (sequence length 378)
 CSP_PLAVI (P08677) CIRCUMSPOROZOITE PROTEIN PRECURSOR (CS).
 Nona-peptide  repeat unit:GDRADGQPA  170/198->
        starting at 96    :GDRADGQPA GDRADGQPA GDRADGQPA GDRAAGQPA GDRADGQPA
                          -GDRADGQPA GDRADGQPA GDRADGQPA GDRAAGQPA GDRAAGQPA
                          -GDRADGQPA GDRAAGQPA GDRADGQPA GDRAAGQPA GDRADGQPA
                          -GDRAAGQPA GDRAAGQPA GDRAAGQPA GDRAAGQPA GNGAGGQAA

WISS-PROT: P23253   (sequence length 1162)
CNA_TRYCR (P23253) SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJO
2-mer repeating unit.....DSSAHSTPSTPA  502/564->
      starting at 593   :DSSAHSTPSTPA DSSAHSTPSTPV DSSAHSTPSTPA DSSAHGTPSTPV
                        -DSSAHGTPSTPA DSSAHGTPSTPV DSSAHSTPSTPV DSSAHSTPSTPV
                        -DSSAHGAPSTPA DSSAHGTPSTPV DSSAHGTPSTPA DSSAHSTPSTPA
                        -DSSAHSTPSTPA DSSAHSTPSTPV DSSAHGTPSTPA DSSAHSTPSTPA
                        -DSSAHGTPSTPV DSSAHSTPSTPV DSSAHGTPSTPV DSSAHSTPSTPV
                        -DSSAHGTPSTPV DSSAHSTPSTPA DSSAHSTPSTPA DSSAHGTPSTPV
                        -DSSAHSTPSTPA DSSAHSTPSTPV DSSAHSTPSTPA DSSAHGTPSTPV
                        -DSSAHGTPSTPA DSSAHSTPSTPA DSSAHSTPSTPA DSSAHSTPSTPV
                        -DSSAHSTPSTPA DSSAHSTPSTPA DSSAHSTPSTPA DSSAHSTPSTPV
                        -DSSAHSTPSTPA DSSAHGTPSTPA DSSAHSTPSTPV DSSAHSTPSTPA
                        -DSSAHGTPSTPA DSSAHSTPSTPA DSSAHGTPSTPA DSSAHSTPSTPA
```

Figure 3.5: A few examples of antigenic proteins containing tandem oligo-peptide repeats

Structural proteins represent another class of proteins containing long oligo-peptide repeats and a few examples can be quoted here. The proline rich plant cell wall structural proteins of *Medicago* and soybean have extensively repeated deca-peptides, PPVYKPPVEK. The cytoskeletal keratin proteins from higher animals contain glycine rich oligo-peptide repeats like GGGL, GGGSF, GGGGF, GGGMGM, and GGFGGA. The skin epidermal keratinocyte proteins, involucrins, loricrins, repetins, and small proline-rich proteins (cornifins) also contain oligo-peptide repeats of various types. The neurofilament-triplet-H proteins of mammalian neuronal axons have tandem hexa-peptide, EAKSPA, repeats where serines are the sites of extensive phosphorylation and cross-linking (Julien and Mushynski, 1998). Other structural proteins containing significant oligo-peptide repeats include hair root cell trichohyalins, tropoelastins, silk moth fibroins, *Drosophila* salivary glue proteins, yeast cell wall proteins, epithelial mucins and cartilage specific aggrecan core proteins.

A classic example of a protein containing evolutionarily conserved oligo-peptide repeats is the largest subunit of RNA-polymerase-II. The carboxy-terminal domain (CTD) of this protein consists of hepta-peptide, YSPTSPS, tandemly repeated for ~6 to 47 times across a wide range of organisms including human, *Drosophila*, yeast, and *Arabidopsis*. The CTD seems to play an important role during transcription activation and also functions as a platform for assembly of multi-protein complexes that hold, splice and poly-adenylate pre-mRNA as it is synthesized by the polymerase (Corden and Patturajan, 1997).

The *Drosophila* Period Circadian Protein contains long stretches of di-peptide glycine-threonine repeats. These repeats are polymorphic in length in geographically distinct populations and are possibly correlated with the ability of flies to maintain a circadian period at different temperatures (Sawyer et al., 1997). The serine-arginine rich splicing factors of human, mouse and chicken have SR di-peptide repeat domains which have been found to be essential for protein-protein interaction and also as splicing activators (Graveley and Maniatis, 1998).

Internal duplications in proteins may be grouped in 3 categories depending on the size of repeating units. In the first case, each of the duplicated domains constitutes structurally and functionally independent unit (e.g. zinc-finger domain, homeo domain, SH2 domain, immunoglobulin domain) and possibly originates from entire exon duplication. The second category duplications are repeats of ~20 to 40 residues that have been identified in several protein families (Groves and Barford, 1999; Andrade et al, 2000; Table 3.6). Crystal structure studies have shown that, in these proteins, each of the repeated motifs adopts distinct structural units and when present in tandem arrays they exhibit striking superhelical structures with characteristic handedness, twist and curvature (Figure 3.6). Each structural unit may be composed of two or more secondary structural elements (e.g. α/α, α/β, β/β, $\alpha/\alpha/\alpha$, $\alpha/\alpha/\beta/\beta$, etc.) and the successive repeating units are stacked through hydrogen-bonding and hydrophobic interactions with the neighbors (Kobe and Kajava, 2000). The main advantage of superhelical structures is that they provide extended surface area and facilitate protein-protein interaction for formation of large protein complexes. Perhaps, this may be the common role played by the helical repeats that otherwise occur in various proteins performing diverse functions.

Table 3.6: Characteristics of some protein sequence repeat families

Name of the repeat	Length of repeat unit	Number of repeats in a protein	Functions of proteins containing repeats
Ankyrin repeat	~33	4-20	Transcription regulation, cytoskeleton organization, developmental regulation, toxins, membrane receptors
Armadillo repeats	40-42	10-12	Cell adhesion, signaling pathway
HEAT repeats	37-43	3-22	Huntington protein, elongation factor, protein phosphatase
Leucine Rich Repeats (LRR)	20-30	4-30	RNAse inhibition, cell adhesion, signal transduction and plant defense
Tetra-Trico-Peptide repeats (TPR)	~34	3-16	Chaperone, cell-cycle, transcription regulation, protein-transport complexes
WD40 repeats	36-46	4-16	G-protein complex, , RNA processing, transcription regulation, cytoskeleton assembly

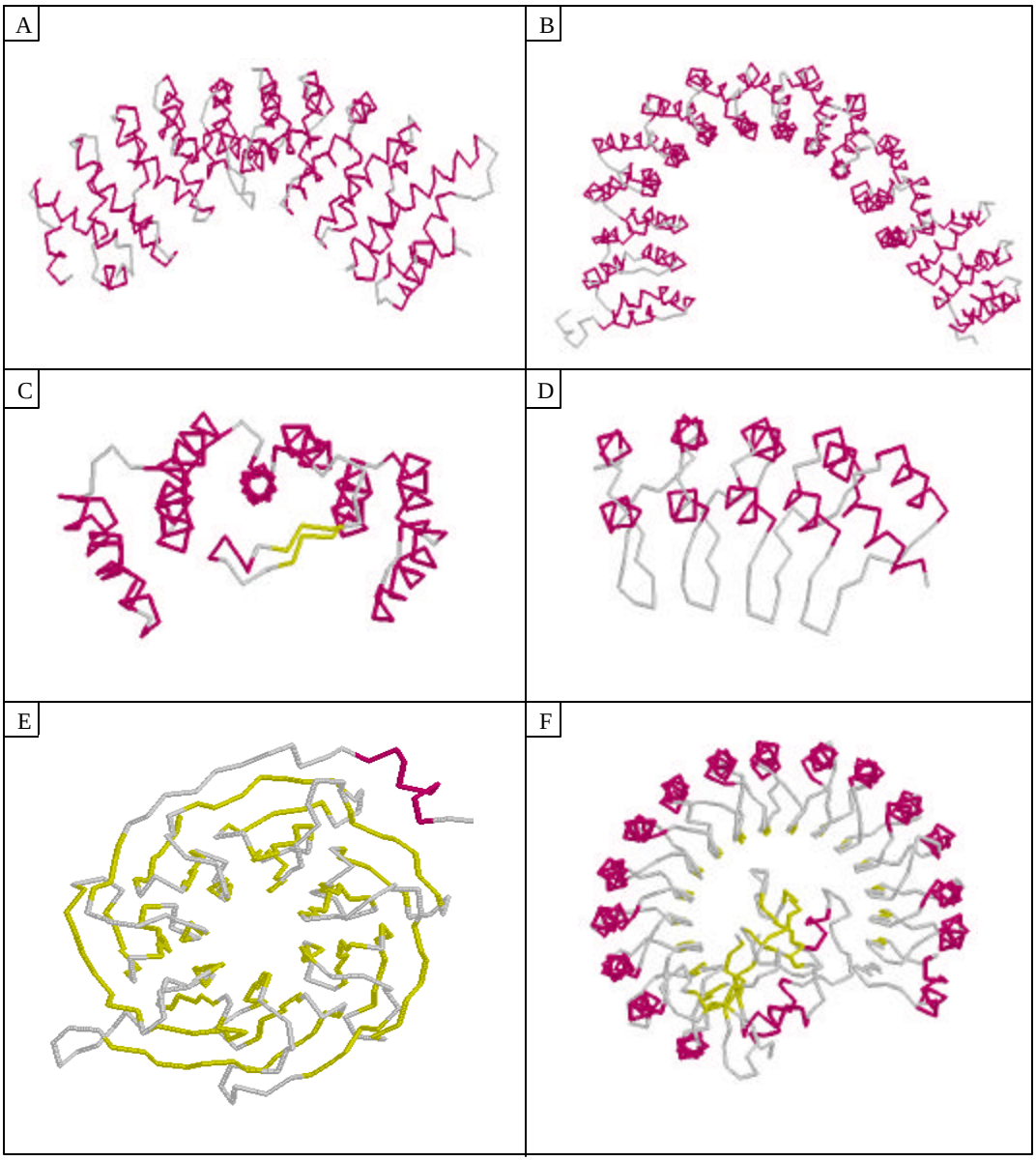


Figure 3.6: Topology of repeated structural arrays observed in some protein sequence repeat families. A> Armadillo repeats (PDB: 1BK5-A); B> HEAT repeats (PDB: 1B3U-A); C> Tetratricopeptide repeats (PDB: 1E96-B); D> Ankyrin repeats (PDB: 1AWC-B); E> WD40 repeats (PDB: 1GP2-B); F> Leucine Rich Repeats (PDB: 1DFJ). All figures were drawn using program RasWin v2.4 (Roger Sayle). Color scheme: Alpha helices as magenta, beta sheets as yellow and all others as gray.

The third category of internal repeats could be tandem repeats of single amino acids or short oligo-peptides. Here, the repeating units are small and are unlikely to form independent structural units. Rather, short oligo-peptide repeating units may promote regular structures when several units appear in succession. For example, crystal structure study demonstrated that tandem imperfect hexa-peptide repeats in UDP-N-acetylglucosamine acetyltranseferase formed left-handed parallel β-helix (Raetz and Roderick, 1995, PDB-ID: 1LXA). Each hexa-peptide unit formed parallel β-strand that resembled a side of an equilateral triangle, which in turn stacked one above the another to form a structure similar to equilateral prism. Bateman et al., (1998) have proposed that similar β-helix structures could be formed by tandem penta-peptide repeats, $[A(D/N)Lxx]_n$, observed in some bacterial proteins. The structural models for bacterial ice-nucleation proteins predict that the consecutive octa-peptide repeat units in ice-nucleation proteins can form parallel-antiparallel β-strands that assemble in 48 residue rectangular units (Kajava and Lindow, 1993; PDB-ID: 1INA). Such rectangular planes present hydrogen bond donors and acceptors in a manner analogous to ice crystal plane and can thus promote ice nucleation.

3.3.3 Periodic conservation of amino acids:

During the course of evolution, tandem oligo-peptide repeats might have undergone substitutions leaving behind only structurally or functionally important amino acids unchanged. Therefore, to detect ancient repeat patterns, I analyzed the protein sequences for amino acids conserved periodically at every second, third, fourth, fifth, sixth, seventh, eighth, ninth or tenth position. From this analysis, several periodic patterns emerged that could not be detected earlier by searching for tandem oligo-peptide repeats. Some of the proteins containing periodic repeats are listed in Table 3.7 whereas, a few interesting patterns revealed from this study are depicted in Figure 3.7.

One of the most striking periodic behaviors is glycine repeated at every third position in collagen proteins, the major structural proteins of bone, cartilage, skin and tendons of higher animals. Three collagen polypeptides wrap around each other to form a triple-helical super-coiled structure which is possible only if glycines occur at every

third position on each chain (Brodsky, 1990). Substitution of a single glycine in type I collagen has been reported to cause misfolding, leading to "brittle bone" disease (Baum and Brodsky, 1999). The collagenic triplet repeats $(GXY)_n$ have been also detected in several globular proteins including collagenic tail peptide of acetylcholine esterase, macrophage scavenger receptor, human complement subcomponent C1q and mammalian c-type lectins like mannan binding protein, lung surfactant protein-D, bovine conglutinin and collectin-43. The collagenic domains in these proteins allow them to trimerize by triple helical winding that facilitates proper functioning of these proteins (Krejci et al., 1997; Kishore and Reid, 1999; Hoppe and Reid, 1994; Andersson and Freeman, 1998).

Periodic conservation of amino acids may be useful in structural packing of two or more polypeptide chains of the same or different proteins. For example, as discussed earlier, glycine at every third position is essential for triple helix formation (Brodsky, 1990). In case of leucine zippers, the leucines conserved at every seventh position fall on a straight line along a side of helix and can zip together with a similar motif of another polypeptide (Landschulz et al., 1988; Figure 3.8). The alternately placed glutamines on two beta-sheets can sterically fit and exchange hydrogen bonds to form polar zippers (Perutz et al., 1994; Figure 3.8). Periodically placed amino acid side chains can also facilitate one to one interactions with target atoms showing similar periodicity. One such example is found in type-I antifreeze protein of winter flounder that contains three $T(X)_2(D/N)(X)_7$ repeats. The regularly placed threonine and aspartate / asparagine residues on this alpha helical protein hydrogen bond with equivalently placed oxygen atoms along <0112> axis of {2021} ice planes and prevent ice crystal growth (Chou, 1992; Sicheri and Yang, 1995; PDB-ID: 1WFA; Figure 3.9).

Table 3.7: Representative proteins containing periodically conserved amino acids

Period	Accession number	Repeating unit	Number of repeats[a]	Organism	Protein name
2	P09789	Gx	160	*P. hybrida*	Glycine-Rich Cell Wall Structural Protein 1
	P05790	Gx	51	*B. mori*	Fibroin Heavy Chain
	P40603	Px	35+20	*B. napus*	Anther-Specific Proline-Rich Protein
	P04265	Gx	43	*X. laevis*	Keratin, Type II Cytoskeletal I
3	P02461	Gxx	352	*H. sapiens*	Collagen Alpha 1(III) Chain
	P05227	Axx	81	*P. falciparum*	Histidine-Rich Protein Precursor
	P35247	Gxx	59	*H. sapiens*	Pulmonary Surfactant-Associated Protein D
	P23805	Gxx	56	*B. taurus*	Conglutinin Precursor
	Q03637	Gxx	55	*T. marmorata*	Acetylcholinesterase Collagenic Tail Peptide
4	P15714	Qxxx	24	*E. tenella*	Antigen LPMC-61
	P49919	Exxx	19	*M. musculus*	Cyclin-Dependent Kinase Inhibitor 1c
5	P53353	Exxxx	42	*V. vulpes*	Sperm Acrosomal Protein FSA-ACR.1
	P36417	Qxxxx	20	*D. discoideum*	G-Box Binding Factor
6	P97347	Qxxxxx	42+12+19	*M. musculus*	Repetin
	Q28824	KPxxxx	41	*B. taurus*	Myosin Light Chain Kinase, Smooth Muscle
	P51861	EDxxxx	34	*H. sapiens*	Cerebellar-Degeneration-Related Antigen 1 (CDR34)
7	P22793	Exxxxxx	35	*O. aries*	Trichohyalin
	Q28983	PTExxxx	33	*S. scrofa*	Zonadhesin
	Q15428	Pxxxxxx	27	*H. sapiens*	Spliceosome Associated Protein 62
8	P16239	SxxxAxxx	131	*E. herbicola*	Ice Nucleation Protein
	P22792	Lxxxxxxx	33	*H. sapiens*	Carboxypeptidase N 83 KD Chain
	P13983	PPxxxxxx	24	*N. tabacum*	Extensin (Cell Wall Glycoprotein)
9	Q28107	QxxLSPDxx	28	*B. taurus*	Coagulation Factor V
10	P14708	QxGQLxxxxx	62	*P. pygmaeus*	Involucrin
	P17437	APAPAxxExx	25	*X. laevis*	Skin Secretory Protein XP2

[a] approximate in some cases since different amino acids within a repeating unit show different levels of conservation

Figure 3.7: A few examples of protein sequences showing periodically conserved amino acids

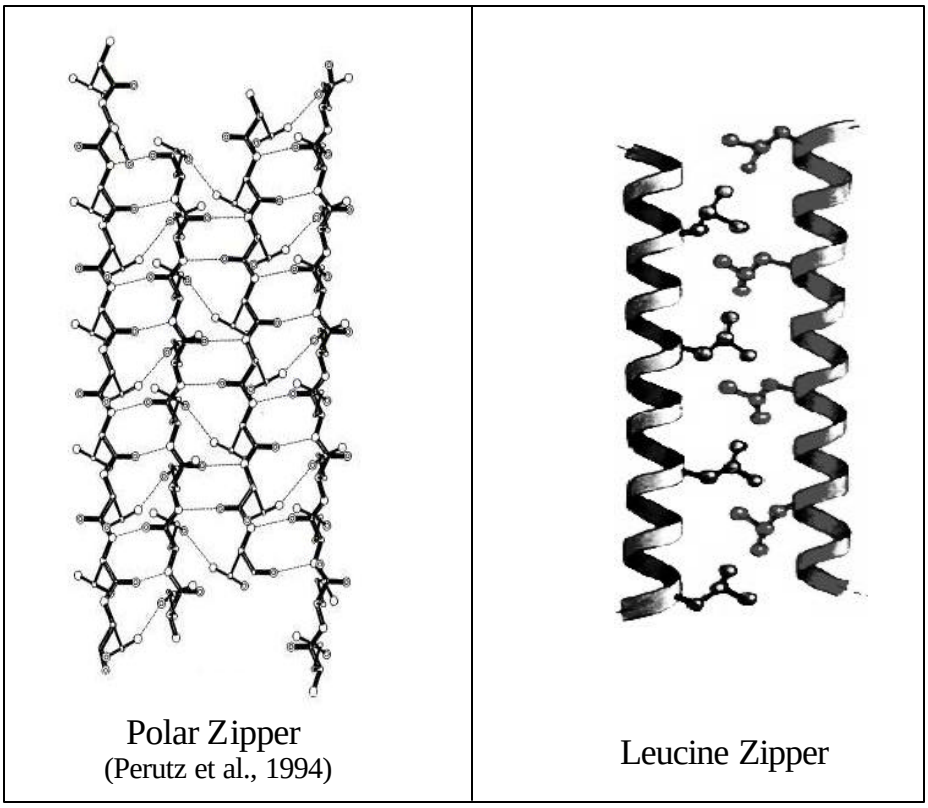


Figure 3.8: Zipper-like interactions shown by periodically conserved amino acid residues

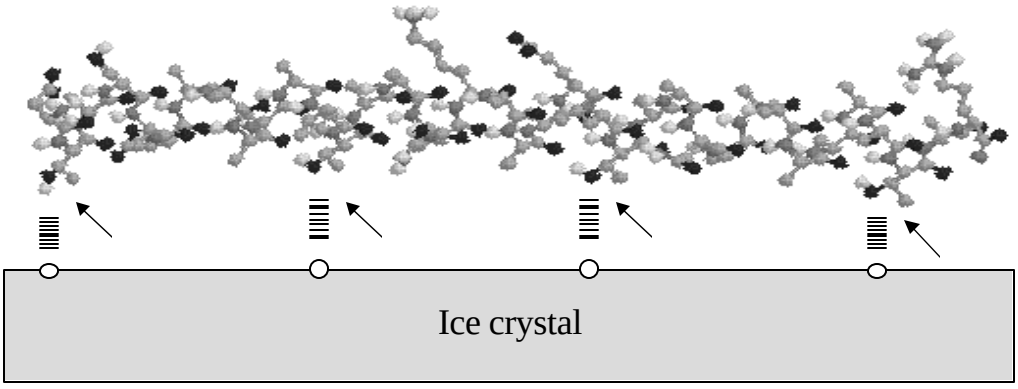


Figure 3.9: Schematic diagram showing binding of anti-freeze protein (PDB-ID: 1WFA) to ice crystal. Arrows indicate conserved threonine residues that hydrogen bond with equivalently placed hydroxyl groups in ice crystal.

3.4 CONCLUSIONS:

My study provides a comprehensive picture of repeat patterns observed in protein sequences. Although, internal repeats have been detected in several proteins and their importance demonstrated in some cases, not much information is available about their exact role in protein structure and function. One advantage of these periodic patterns is that they juxtapose similar functional groups in space and thereby facilitate zipper-like interactions with target molecules. This provides a different perspective for prediction of structural models and design of novel proteins. We hope that the extent of repeat patterns as revealed from our database will be useful for further analysis of internal repeats with respect to their origin, evolution and their implications on protein structure and function.

ABSTRACT

Comparative promoter analysis is a promising strategy to identify putative regulatory motifs conserved in evolutionarily related sequences or in genes showing common expression profiles. To facilitate such analysis, I have developed a software tool that detects conserved transcription factor binding sites, cis-elements, palindromes and k-tuples simultaneously in a set of promoter sequences. When promoter sequences of diverse members of an orthologous gene family are analyzed, the evolutionarily conserved motifs can be identified and such sites can be expected to have a functional role. The program developed by me can also be used to study promoters of genes showing co-ordinate patterns of expression to check if they have similar regulatory modules. Information from such analysis can be useful in understanding modular organization of promoters and designing further experiments to unravel genomic cis-regulatory logic programmed in DNA sequences. The program TRES has been implemented on a web-server and can be used from the URL: http://bioportal.bic.nus.edu.sg/tres

4.1 INTRODUCTION:

Among all the genes encoded in a genome, only a few are expressed at a particular time in a particular tissue. For proper growth, development and survival of an organism it is essential that specific proteins or gene products be synthesized in appropriate amount at appropriate time and space. The genes are complete information units in the sense that they not only code for proteins but also contain address label (promoter) that specifies where and when each of them should express. A typical promoter is an array of specific modules (short DNA sequences) separated by strings of non-specific bases and organized sequentially around transcription initiation site (Maniatis et al., 1987). Different transcription factors bind to these modules in a sequence specific manner and by cooperative interaction they bring about favorable changes in local chromatin structure and participate in assembly and activation of transcription initiation complex (Ptashne, 1988; Buratowski, 1994; Tjian and Maniatis, 1994; Carey, 1998; Kadonaga, 1998).

To identify sequence motifs involved in the transcriptional regulation of a gene, one approach is to search for known transcription factor (TF) binding sites in its promoter DNA sequence and then design experiments to verify if conserved putative motifs play any role in the regulation of gene expression. Once a putative region of a promoter DNA sequence is identified it can be mutated, fused to a reporter gene and its effect on gene expression level can be studied by transformation. During the last two decades, a large number of transcription factor binding sites, cis-elements and enhancer elements involved in the regulation of various genes from diverse organisms have been identified and characterised. Databases such as TRANSFAC (Heinemeyer et al., 1999), ooTFD (Ghosh, 2000) and PLACE (Higo et al., 1999) provide an updated compilation of these elements. Several computational tools have been developed that search for putative regulatory sequence motifs in a given promoter sequence (e.g. SIGNAL-SCAN (Prestridge, 1996), ConsInspector (Frech et al., 1997d)). However, it is feared that string searches based on IUPAC consensus sequence do not consider certain allowed mismatches and differential importance of bases frequently observed in protein-DNA interactions (Stormo and Fields, 1998; Frech et al., 1997b). Therefore, an alternative strategy that is more widely used is to search based on the position weight matrices calculated by considering frequency of

each base at each position in a motif. The matrix based search is more reliable and also predicts the strength of a motif in a given promoter sequence (Chen et al., 1995; Quandt et al., 1995; Frech et al., 1997c).

When a promoter DNA sequence is searched for putative TF binding sites from large databases (e.g. TRANSFAC), it is often noticed that several motifs appear to be conserved all over the promoter sequence. However, not all of them could be expected to be involved in transcription regulation and some of the motifs might occur by chance alone in a sequence, thus making it difficult to choose for further experimental analysis. It is suggested that phylogenetic conservation of regulatory motifs in sufficiently diverse orthologous genes can provide a rigorous testimony of their functional role (Duret and Bucher, 1997). This "phylogenetic footprinting" approach has been used to identify evolutionarily conserved regulatory modules in globin genes (Gumucio et al., 1996) and in light responsive plant promoters (Arguello-Astorga and Herrera-Estrell, 1996). More recently, Mironov et al., (1999) have demonstrated usefulness of comparative promoter analysis in their study of orthologous regulons from *E. coli* and *H. influenzae*.

Although comparative promoter analysis is more informative to identify putative regulatory elements, existing programs do not allow simultaneous analysis of related sequences. Therefore, I designed a computer program that would search for known TF binding sites, palindromes and highly conserved k-tuples simultaneously in a set of sequences. This program has also been implemented on a web-server so that it can be easily used through the Internet.

4.2 MATERIALS AND METHODS:

4.2.1 Program organization:

The program TRES (Transcription Regulatory Element Search) is written in "C" and implemented on Unix server. Using TRES, as many as 20 promoter sequences, each of maximum 1000 bp length, can be simultaneously searched for putative regulatory elements. TRES has been organised in following 4 analysis tools:

4.2.1.1. Matrix-search: This program scans the input sequences for conserved TF binding sites using matrices described in TRANSFAC database (Heinemeyer et al., 1999). From the nucleotide frequency distribution matrices, the position weights and matrix similarity scores are calculated essentially according to Quandt et al., (1995) except that gaps are not considered and a pre-processed library of normalised weight matrices is used during runtime. For a particular nucleotide distribution matrix the position weights (po_wt(i)) for each position(i) are calculated as,

$$po\_wt(i) = (100/\ln(4)) \times \Big(\sum_{b \in \{A,T,G,C\}} [\, rbf(b,i) \times \ln(rbf(b,i)) \,] + \ln(4) \Big)$$

where *rbf*(b,i) is relative base frequency of base(b) at position(i). For any matrix, maximum score (matrix_max_score) is calculated as,

$$matrix\_max\_score = \sum_{i=1 \text{ to } n} (po\_wt(i) \times max\_rbf(i))$$

Where, (n) is length of the matrix and max_*rbf*(i) is maximum relative base frequency at position(i). In order to avoid recalculations during each runtime, a pre-processed library of normalised weight matrices has been created using a 'C' program. The normalised weights for each base(b) at each position(i) are calculated as,

$$nomalised\_wt(b,i) = 100 \times (po\_wt(i) \times rbf(b,i)) / (matrix\_max\_score)$$

During matrix scanning, sliding along each sequence, the matrix similarity score is calculated by simply adding normalised weights.

$$matrix\_similarity\_score = \sum_{i=1 \text{ to } n} (normalised\_wt(b,i))$$

This directly gives a comparative value in the range of 0 to 100. A TF binding site is considered to be conserved only if the matrix similarity score falls above the user defined cut-off value in the range 75 - 100. A representative nucleotide frequency distribution matrix and calculated normalised weights are shown in Table 4.1.

Table 4.1: Nucleotide frequency distribution matrix* (TRANSFAC Acc. No. M00123) and calculated normalised weights for the occurrence of different nucleotides at different positions in c-Myc/Max transcription factor binding motif.

Position in the motif	1	2	3	4	5	6	7	8	9	10	11	12
Nucleotide frequency:												
A	7	21	3	0	29	0	9	2	0	7	4	14
C	7	3	11	29	0	27	0	4	0	3	12	4
G	5	1	9	0	0	0	20	0	27	6	7	0
T	10	4	6	0	0	2	0	23	2	13	6	10
position weight	2.7	38.1	6.8	100	100	81.9	55.3	53.7	81.9	8.87	5.69	28.4
Consensus sequence	N	A	N	C	A	C	G	T	G	N	N	W
Normalised weights:												
A	0.11	5.7	0.15	0	20.6	0	3.55	0.77	0	0.44	0.16	2.83
C	0.11	0.81	0.54	20.6	0	15.7	0	1.53	0	0.19	0.49	0.81
G	0.08	0.27	0.44	0	0	0	7.88	0	15.7	0.38	0.28	0
T	0.15	1.09	0.29	0	0	1.17	0	8.8	1.17	0.82	0.24	2.02

* Frequency of occurrence of different nucleotides at different position in the motif were determined by Blackwell et al., (1993) using *in vitro* binding site selection assay.

4.2.1.2. IUPAC-string search: Using this program, input sequences can be searched for TF binding sites or cis-acting elements based on IUPAC consensus sequences described for the sites. Currently, a total of 3980 TF binding sites from TRANSFAC database (Heinemeyer et al., 1999), 5919 sites from ooTFD database (Ghosh, 2000) and 240 plant cis-acting elements from PLACE database (Higo et al., 1999) can be searched.

4.2.1.3. Palindrome search: This tool detects different palindromic sequences (perfect, as well as a few odd base(s) included) in the sequences. If $b_i \in$ {A, T, G, C}, N is any base and c_i is complementary base to b_i, then the program searches for following palindromes:

a. Tetrameric palindromes of the form $b_1b_2(0\text{-}5N)c_2c_1$

b. Hexameric palindromes of the form $b_1b_2b_3(0\text{-}5N)c_3c_2c_1$.

c. Octameric palindromes of the form $b_1b_2b_3b_4(0\text{-}5N)\,c_4c_3c_2c_1$.

d. Decameric palindromes of the form $b_1b_2b_3b_4b_5(0\text{-}5N)c_5c_4c_3c_2c_1$.

4.2.1.4. k-tuple search: This program searches for any string of length 5 to 50 bases conserved in all the sequences. If size of a string is k, instead of searching all 4^k possible words, the program searches for only the subset that is represented in all the sequences under study. TRES searches for individual k-tuple string $<\boldsymbol{base}_{ij}\ldots\ldots\boldsymbol{base}_{i(j+k-1)}>$ where i = 1 to n (number of sequences) and j = 1 to ($seq\_length_i - k$). Essentially, each k-tuple is a window of size 'k' sliding over all the sequences and searched on both the strands at a given mismatch level.

4.2.2 Search and report parameters:

a. Mismatch level: A user of the program can select mismatch level that can be tolerated to consider a match. User can specify either no mismatch or a maximum of 1 mismatch for every 15, 12, 10, 8, 7 or 6 bases of recognition sequence.

b. The location of the sites can be obtained with respect to TATA-box or beginning of the sequence or end of the sequence. The first option is activated only if the TATA-box has been detected initially using a weight matrix (Bucher, 1990), by convention in last 150 bases of all the sequences.

c. The sites can be reported only if they are conserved in all the sequences or if present in a minimum user defined number of sequences.

4.3 RESULTS:

4.3.1 Implementation of TRES on the web:

The TRES program is implemented on Unix server at the URL: http://bioportal.bic.nus.edu.sg/tres. This URL provides a web form (Figure 4.1) wherein users can cut and paste their sequences in a text box, select one of the 4 program modules and associated parameters and submit their sequences for online search. At the server, a "CGI Perl script" is invoked that receives the information and

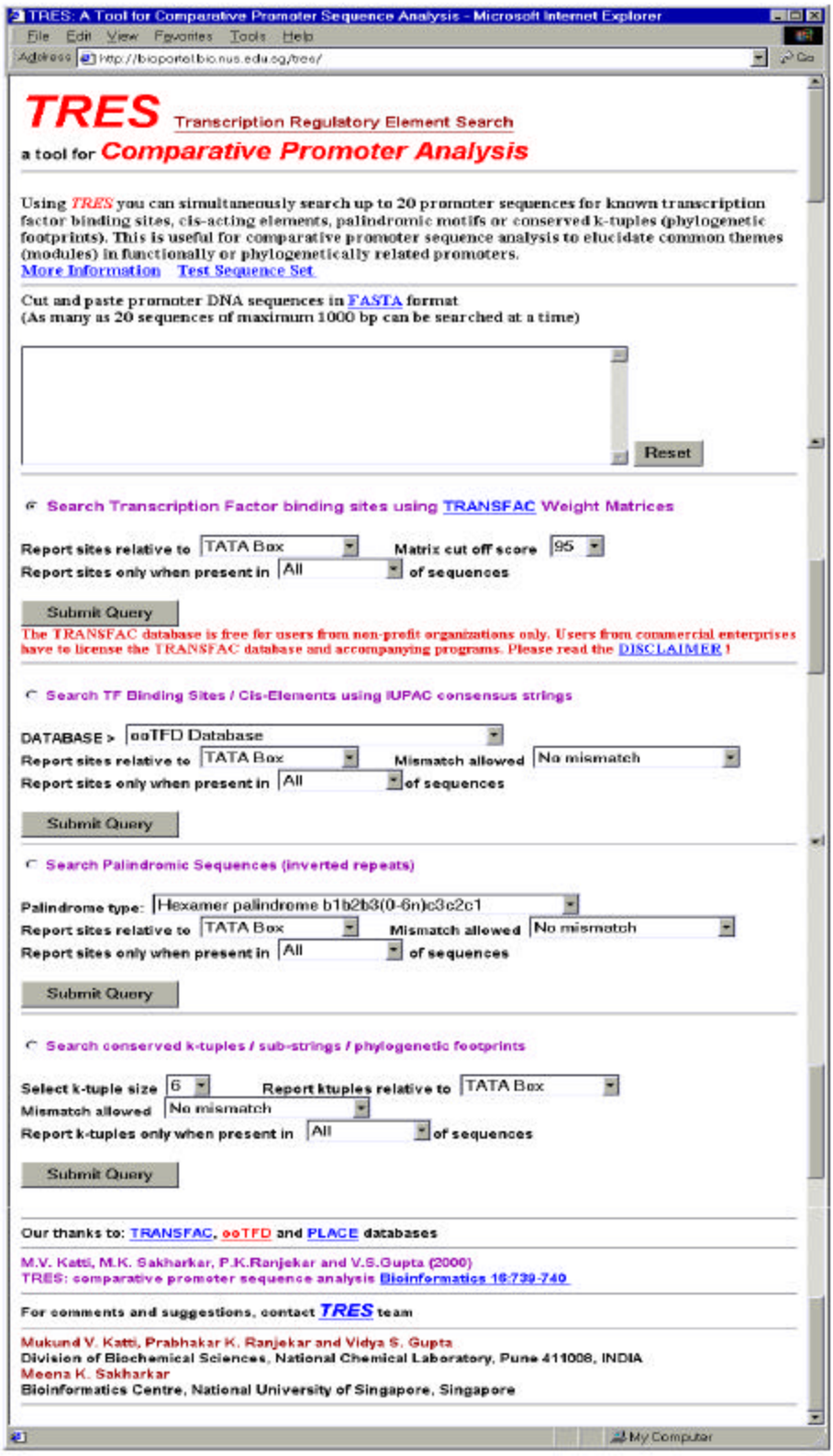


Figure 4.1: Web interface for the program TRES, available at the URL: **http://bioportal.bic.nus.edu.sg/tres**

passes on to the "C" program that analyses the sequences and sends back the results to the client. For all the TRANSFAC, ooTFD and PLACE sites reported in the results, a hyperlink is provided to the corresponding entry in the respective database and thereby further information can be explored. Additional information on the use of TRES and test sequence set is also available on the web.

4.3.2 Application of the program:

The application of program TRES has been exemplified using a set of αA-crystallin gene promoter DNA sequences from human, mouse, mole rat and chicken. Crystallins are the structural proteins specifically expressed in vertebrate eye lens and constitute ~80 to 90% of total lens soluble proteins imparting transparency and optimal refractive index to lens (Kantorow et al., 1993). The regulation of αA-crystallin gene expression is highly specific and limited to lens epithelia and fiber cells. Promoter deletion experiments in cultured cells have shown that about -111 to +46 region of mouse αA-crystallin promoter or -162 to +44 region of chicken αA-crystallin promoter sequence is sufficient for lens specific expression (Ilagan et al., 1999). Some of the important regions involved in the regulation of αA-crystallin genes are depicted in Figure 4.2.

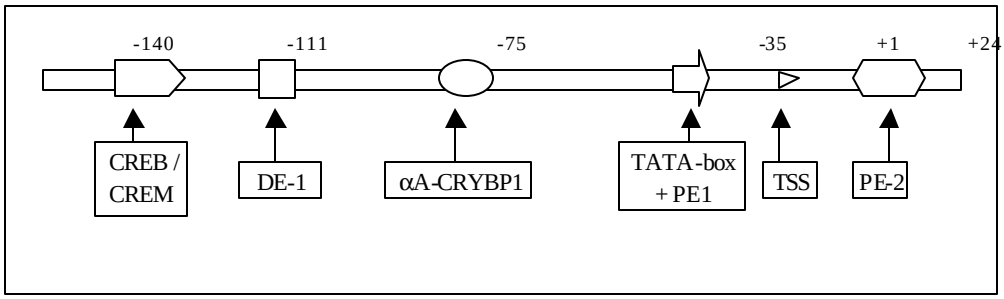


Figure 4.2: A schematic diagram showing regulatory elements involved in αA-crystallin gene expression. DE- distal element; PE- proximal element; TSS- transcription start site.

In my analysis, I used ~400 bp upstream promoter sequences of human, mouse, mole rat and chicken αA-crystallin genes. The pairwise alignment scores between these sequences range from 42 to 70 (Clustal-W, Thompson et al., 1994). Figure 4.3 shows part of the output files obtained from TRES TRANSFAC site search and k-tuple search. The program could detect all the known regulatory elements (Figure 4.2) implicated in regulation of αA-crystallin genes. One of the novel observations from my study is detection of a 8-tuple (TGGGGCTG) conserved at about -110 relative to TATA box in all the sequences (Figure 4.3c). This region could not be detected by Clustal-W multiple alignment program since it is conserved on positive strand in human, mouse and mole rat sequences whereas it is present on both positive and negative strand at different locations in chicken promoter sequence (Figure 4.4). Interestingly, this region corresponds to a putative USF binding-site-A known to be important in regulation of chicken αA-crystallin genes (Cvekl et al., 1994). However, it has not been noticed or characterised in mouse αA-crystallin promoter sequence (Ilagan et al., 1999) probably because of its altered location.

4.4 DISCUSSION:

4.4.1 Salient features of program *TRES*:

Transcription factors or their DNA binding domains are known to be conserved across a wide range of evolutionarily diverse families. Many transcription factors regulate diverse set of genes and each gene may require complex assemblage of various transcription factors for its activation. Therefore, known transcription factor binding sites are the potential regulatory elements to search in new genes / promoters under study. In order to detect TF binding sites in DNA sequences, I have included both the matrix based and IUPAC consensus string based searches in TRES. For the matrix search, the nucleotide frequency distribution matrices from the TRANSFAC database (Heinemeyer et al., 1999) have been pre-computed into normalised weight matrices, which helps in saving the runtime. The IUPAC string based search uses consensus sites described in TRANSFAC (Heinemeyer et al., 1999), ooTFD (Ghosh, 2000) and PLACE (Higo et al., 1999) databases. Some of the disadvantages of IUPAC string search, compared to matrix search, are compensated since user defined mismatches can be tolerated and possibility of false positives can be reduced if many related sequences are searched simultaneously.

A

```
TRANSFAC site name.....     R02115 AACRYBP1$CONS
Site consensus sequence     GGGAAATCCC.

S79457 Human                   -36 *
S79462 Mouse                   -35
M17247 Mole Rat                -37 *
M17627 Chicken                 -39 *
```

B

```
10-tuple-> TGCTGCTGAC (compli. sequence GTCAGCAGCA)

S79457 Human                   -80 TTCTGCTGAC*
S79462 Mouse                   -82 AGCTGCTGAC*
M17247 Mole Rat                -84 TGCTGCTGAC
M17627 Chicken                 -83 TTCTGCTGAC*
```

C

```
8-tuple->> TGGGGCTG  ( compli. sequence..  CAGCCCCA )

S79457 Human:                  -114 TGGGGCTG
   on-compli-strand->          -180 TGGGGCTC*
S79462 Mouse:                  -110 TGGGGCTG
M17247 Mole-Rat                -111 TGGGGCTG
M17627 Chicken:                -130 TGGCGCTG*
   on-compli-strand->          -106 TGGGGATG*
```

Figure 4.3: Part of TRES output files showing (A) AACRYBP1 site (TRANSFAC site search) (B) a 10-tuple corresponding to DE-1 element and (C) a 8-tuple corresponding to USF binding site A (USF-bsA), conserved in human, mouse, mole rat and chicken αA-crystallin promoter sequences. Site locations are relative to TATA box and * indicates one mismatch.

```
S79457_Human        CCAGGCCTCGGGGACAGTCCGTGCACGCTCCTGGGGCTGGG--GGCGGGCACTTGTCCCA  -88
S79462_Mouse        ---AGCCTCTGCTGC--TCAGCGTGTGTTGCTGGGGCTGGGCAGGCGGGTGAGCATTCCA  -82
M17247_Mole rat     ACTGGCCCCTGCTGC--TCAGTGTGTGTCCCTGGGGCTGGGGTGGAGGGTAAG-ATCCCT  -84
M17627_Chicken      TCTGGCGCTGGTTCCCACCAGACTGTCATCCCCAGGTCAGT-------CTCCGCATTTCT  -80
                        **    *   *   * *         *   **   *              *  *
```

Figure 4.4: Part of the multiple alignment of human, mouse, mole rat and chicken αA-crystallin promoter sequences showing a conserved 8-tuple, corresponding to USF-bsA (Figure 4.3c), undetected by alignment program.

Another class of potential transcription regulatory elements are palindromic sequences that show unique features of dyad symmetry and the ability to form hairpins or loops, facilitating protein binding in homo- or hetero-dimer form. For example, the ***b-zip*** and ***b-HLH*** family of plant transcription factors identify core palindromic sequence ACGT and CANNTG, respectively and they bind as homo- or hetero-dimer (Meshi and Iwabuchi, 1995). Important advantages of dimerisation are stability of protein-protein and protein-DNA interactions and generation of diversity from a limited number of transcription factors (Lamb and McKnight, 1991). Therefore, conserved palindromes are strong candidates to be considered as potential transcription regulatory elements. TRES provides a convenient tool to detect different types of palindromic motifs conserved in a set of promoter sequences.

The k-tuple search is useful to identify significantly conserved words in a set of sequences. TRES is more powerful than multiple alignment programs, particularly if conserved words are located at different positions / strands in different sequences or contain a few mismatches. For example, TRES detected a 8-tuple, TGGGGCTG, conserved in different αA-crystallin promoter sequences (Figure 4.3c) which was not obvious in multiple alignment (Figure 4.4). Wolfertstetter et al., (1996) have described a tuple search program to identify functional elements from a set of unaligned sequences based on maximisation of information content. In contrast to their complex algorithm, I have used a simple sliding window method to detect conserved k-tuples of user defined size. It should be noted that if the sequences are highly similar, a very large number of conserved k-tuples are detected. My study suggests that the sequences with similarity score in the range of ~40 to 60% provide a considerable noisy background for k-tuple search.

TRES is useful to study conservation of TF binding sites relative to TATA box. In majority of RNA polymerase-II promoters, TATA box is the site of assembly of transcription machinery and provides a useful reference position for accurate initiation of transcription. Different transcription factors that appear during the course of development or those activated by specific environmental stimulus bind to DNA in a sequence specific manner and interact directly or indirectly in assembly and activation of transcription initiation complex (Ptashne, 1988; Buratowski, 1994; Tjian and Maniatis, 1994; Kadonaga, 1998). For proper protein-protein interaction, the TFs

bound to DNA must approach close enough in proper orientation that may be facilitated by proximity of their cognate binding sites or by looping out of the intervening DNA helix. The composite response elements, which bring TFs in close proximity, facilitate unique combinations of functionally redundant TFs thus generating novel patterns of regulation (Miner and Yamamoto, 1991). For such composite response elements, there might be constraints on spacing between TF binding sites (Kel et al., 1995; Ficket, 1996). However, DNA being a highly dynamic polymer that can bend, twist, roll, stretch, slide, wind and unwind there can be considerable flexibility particularly if long-range protein-protein interactions are involved. Nonetheless, spatial conservation of regulatory motifs relative to TATA box indicates that transcription factors binding to such sites might be directly interacting with the initiation complex assembled at TATA box.

4.4.2 *TRES* is useful to study phylogenetically or functionally related promoter sequences:

When a single promoter sequence is searched for putative regulatory elements, a conserved motif may occur by chance alone in a sequence. On the other hand, if a motif is searched in a set of promoter sequences, the probability of its random occurrence simultaneously in all the sequences is less. Besides, comparative sequence analysis also gives important clues about the spatial organisation of different motifs in context to each other. Therefore, it is obvious that conservation of a motif in a set of sequences is more significant and informative than its detection in a single sequence provided that sequences in the set are not too similar.

Each time a DNA sequence is replicated, there are chances of mutation due to rare failure of proof reading by DNA polymerase or by replication slippage or by misreading of a base that has undergone chemical change. Mutations that do not interfere with the normal functioning continue to accumulate whereas if they affect the vitality, they are selected against during the course of evolution. Therefore, phylogenetic conservation of a sequence motif in an otherwise noisy background strongly suggests its functional role (Duret and Bucher, 1997; Hardison, 2000).

To identify evolutionarily conserved functional motifs, sequences from moderately diverse species should be selected so that there has been sufficient evolutionary time for mutations to accumulate in non-functional regions. It has been suggested that orthologous genes from species with cumulative phylogenetic branch lengths greater than ~200 million years are good candidates for such comparative analysis (Duret and Bucher, 1997; Gumucio et al., 1996). Phylogenetic footprints have been defined as six or more contiguous conserved bases in multiple alignments of orthologous sequences. Present phylogenetic footprint analysis techniques use multiple alignment programs to detect conserved regions in a set of sequences (Gumucio et al, 1996). However, functional modules that are reshuffled or that have undergone substitutions may not be identified by multiple alignment. Since k-tuple search can detect the motifs conserved anywhere in the sequences or on either of the strands, I suggest that k-tuple search is a powerful tool for phylogenetic footprint analysis.

TRES is also useful to study functionally related promoter sequences. If two or more genes are expressed co-ordinately in the same tissue at same time or in response to same environmental stimulation, the questions arise, whether such genes are regulated by same mechanisms and whether similar kinds of transcription factors are required for their activation? An insight into answers to these questions can be obtained by studying conservation of potential regulatory elements in functional promoter regions of genes that show similar patterns of expression. Such an approach has been used to identify regulatory elements conserved in a set of muscle specific genes (Wasserman and Fickett, 1998). Frech et al., (1997a) have developed a method to generate regulatory model from a set of sequences based on modular nature of promoters. Our program uses similar modular approach and can be useful to generate initial model for further use of program *Model Generator*.

Recent developments in micro-array based mRNA quantification make it possible to identify a large number of genes with common regulatory programs (Bucher, 1999). Such set of genes can be used to identify common regulatory modules involved in their expression. For example, Harmer et al., (2000) examined temporal patterns of gene expression in *Arabidopsis* plants using GeneChip arrays representing ~8200 different genes and observed that ~6% of the genes exhibited circadian changes in the steady-state mRNA levels. Further, comparative analysis of upstream promoter

sequences revealed a conserved "evening element", AAAATATCT possibly involved in conferring circadian rhythmicity in plants. Thus, with ever-increasing availability of sequences and their expression profiles, comparative promoter analysis appears to be a promising strategy to identify regulatory modules in genes of interest (Figure 4.5).

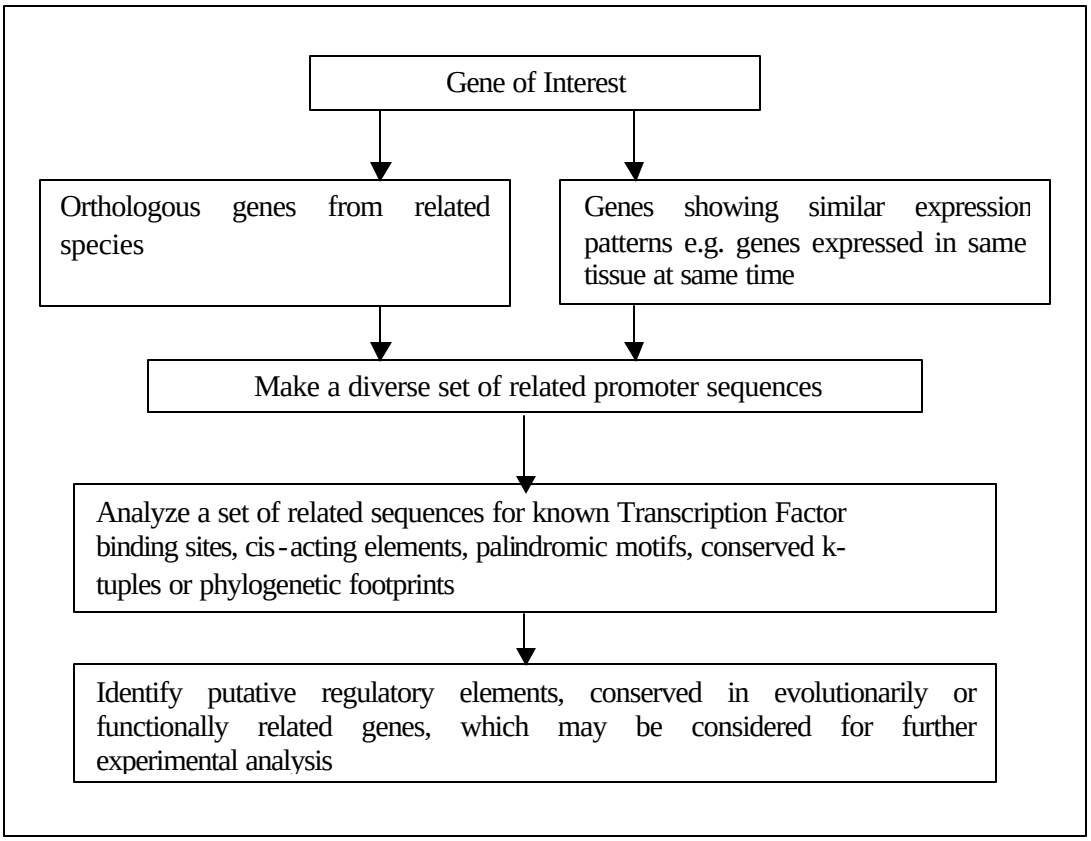


Figure 4.5: A strategy for comparative promoter sequence analysis

4.5 CONCLUSIONS:

Our program TRES provides a useful tool to analyse conservation and spatial distribution of potential transcription regulatory elements simultaneously in a set of sequences. TRES makes use of known information on transcription factor binding sites / cis-elements and at the same time can detect new putative motifs (palindromes, k-tuples or phylogenetic footprints). The main advantage of TRES over other available programs is that it can analyse many related sequences at a time and report only the sites that are conserved in all or in majority of the sequences. Thus, motifs that occur only in one or a few sequences, possibly due to chance, can be filtered.

I conclude that instead of searching for potential regulatory elements in a single promoter sequence, it is more informative to search simultaneously in a set of functionally or phylogenetically related promoter sequences. Though our program is not aimed to predict any models *per se*, it helps to identify potential regulatory modules which researchers can consider in context of available knowledge, develop their own models and design further experiments. As shown by Yuh et al., (1998), by carefully designed experiments coupled with computational analysis, it is possible to unravel genomic cis-regulatory logic programmed in DNA sequences.

THESIS OVERVIEW:

At the dawn of the new millennium, bio-medical research has arrived at a very exciting stage. One of the important paradigm shifts at this juncture has been, unraveling of the complete genome DNA sequence information of human and several other eukaryotic and prokaryotic organisms. However, although this provides us information about total sets of genes that shape a living entity, functions of a large number of predicted genes still remain unclear and determining functions of these genes will be the major task for a next few years. The knowledge of complete genome sequence will help us in designing experiments to elucidate complex metabolic networks and understand cascading of genes during development. Further, from the human genome sequence information it will be possible to map exact causal genes involved in hereditary disorders and susceptibility to disease and this will help in development of preventive medicine and new therapeutic approaches.

The exponential growth of biomolecular sequence data has necessitated development of automated tools to retrieve meaningful information from the raw sequence data. This has led to the advent of a new science of Bioinformatics that acts as an interface between biology, mathematics, computer science and information technology. The pursuit of Bioinformatics is not only to manage and disseminate biological data, but is also to develop new tools and analyze the information to discover new facts, relationships and biological principles. Some of the grand challenges for Bioinformatics for the next few years will be protein structure prediction, finding significant sequence homologies particularly in twilight zones, phylogeny construction and genome sequence analysis (Searls, 1998). The computational genomics will have to address a more integrated analysis of genome information to understand metabolic pathways, signaling networks, functional grouping, phylogenetic patterns and protein fold types (Tsoka and Ouzounis, 2000).

In my thesis, I have attempted to analyze complete genome/chromosome sequences available from a few eukaryotic species, to have an insight into the organization of simple sequence repeats at whole genome/chromosome level. My study reveals that different genomes show characteristic distributions of various repeats and the

abundance or rarity of different repeats in a genome can not be explained by nucleotide composition of a sequence or potential of repeated motifs to form alternative DNA structures. These observations have implications on current theories explaining genesis of repeats where DNA-strand slippage mediated errors during DNA replication and repair have been considered to be the major mechanism in expansion or deletion of repeat tracks. Alternative DNA structures formed by some of the repeated motifs are thought to stabilize strand slippage and expedite slippage events. However, since different genomes show different trends in enrichment of specific repeats, it appears that apart from nucleotide composition of repeat motifs, species-specific cellular factors interacting with them are also likely to have an important role in the genesis of repeats.

Several researchers have used *E. coli* and yeast model systems to study mechanisms of repeat expansion. Indeed, these studies have provided valuable information about relationship between instability of microsatellite loci and cellular factors involved in DNA replication, recombination and mismatch-repair. However, considering the characteristic differences in microsatellite distributions in human, compared to yeast or *E. coli*, it becomes obvious that we are still far from a model system to understand abnormal repeat expansions involved in several human neurodegenerative disorders. Availability of a large number of microsatellite loci identified from complete chromosome sequences should now allow direct investigations to understand the location and sequence dependent instability of microsatellite loci in different genomes.

Genesis of simple sequence repeats could be considered as an aberration in normal DNA processing and these aberrations can occur even in the protein coding regions of DNA, leading to appearance of repeated sequence patterns in proteins. Since simple sequence repeats mutate by additions or deletions of whole repeating units, these events may alter the reading frame and can drastically change the amino acid sequence of a protein. For example, when a dinucleotide repeat in a protein coding region expands by addition of one dinucleotide unit the reading frame would be altered by +2 beyond 3' end of the repeat track. However, deletions or expansions of trinucleotide repeats or multiples thereof (e.g. hexanucleotide repeats) do not alter the reading frame since nucleotides are added or removed in multiples of three.

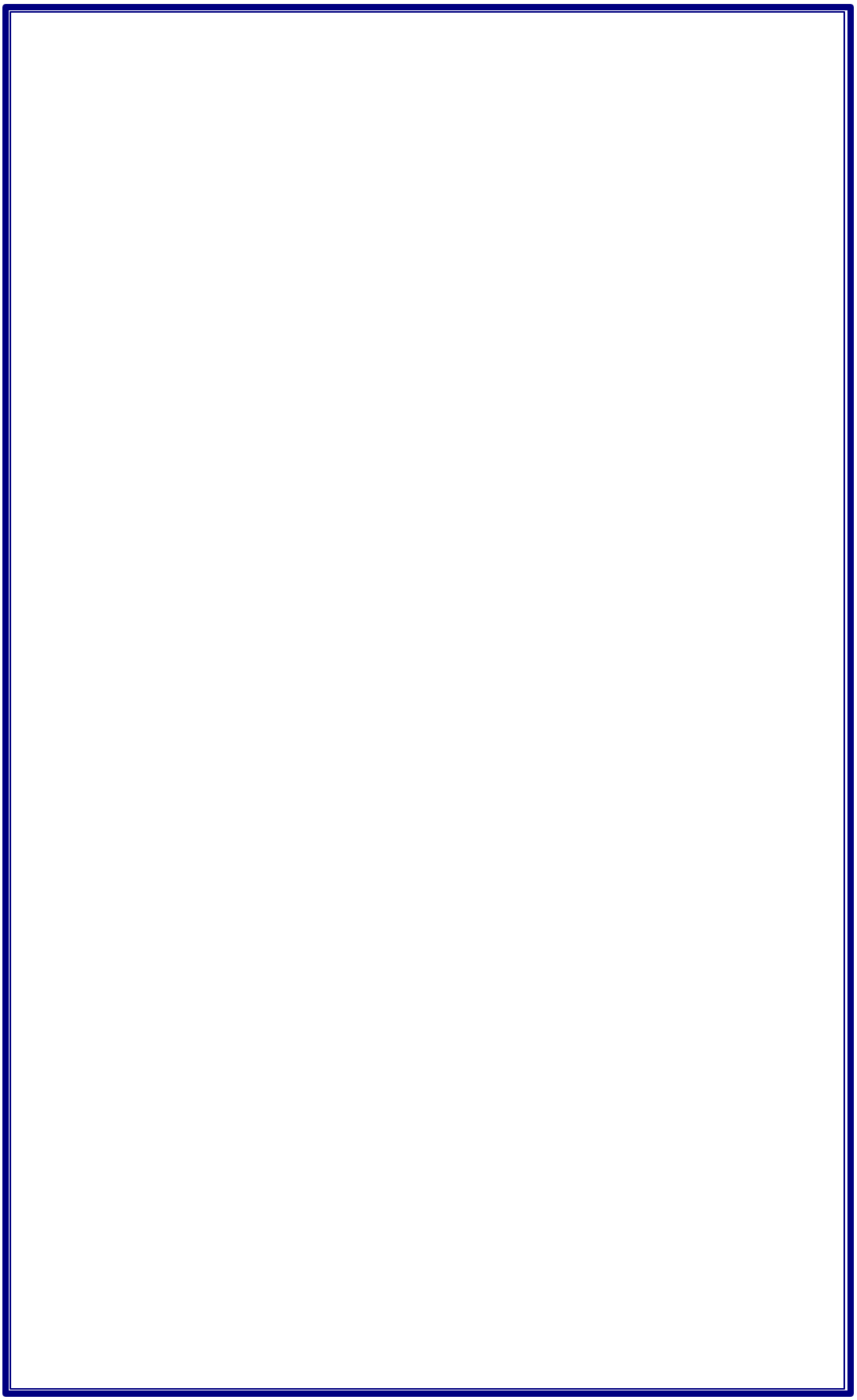
Emergence of a trinucleotide repeat in a coding sequence can result in appearance of a single amino acid repeat stretch in the encoded protein. However, successive expansion events in a trinucleotide repeat region can cause further expansions in single amino acid repeat region of the protein that, beyond certain limit, can drastically affect protein structure and function. From the analysis of complete genome coding DNA sequence sets of yeast, *C. elegans* and *Drosophila*, I have found that expansions of codon repeats corresponding to small hydrophilic amino acids are more tolerated compared to codon repeats encoding hydrophobic amino acids. These observations were further substantiated from the analysis of all the protein sequences from the SWISS-PROT database. Perhaps, expansions of single amino acid repeats of small hydrophilic amino acids are likely to be tolerated if they occur in the linker regions and if they can be easily solvated on surface of the proteins. On the other hand, expanding stretches of hydrophobic amino acids probably collapse towards interior causing protein misfolding.

In addition to single amino acid repeats, I have also studied occurrences of short tandem repeats in protein sequences and have observed that internal repeats of various types, lengths and sequences occur in several proteins. Since amino acid sequence of a protein determines its structure, it would be interesting to know whether repeated sequence patterns are reflected in repeated structural patterns. One advantage of these repeated patterns could be that they can provide regular arrays of spatial and functional groups that could be useful for structural packing or for one to one interactions with target molecules. Indeed, researchers have identified several protein families containing internal repeats where each of the repeating units forms a distinct structural unit. However, majority of these families have longer repeating units (~20 or more residues) and we still do not know much about the structures formed by short tandem repeats. We hope that the wide range of internal repeats observed in protein sequences, as revealed from our database (TRIPS), will bring greater interest among researchers to undertake further studies in this direction.

Comparative sequence analysis is a very informative approach to elucidate evolutionary relationships and to understand structurally and functionally important regions in the DNA and proteins. Similarly, comparative analysis of related promoter sequences is increasingly being considered as an effective strategy to identify

functional regulatory modules. I have designed a computer program, TRES, which allows simultaneous analysis of several promoter sequences to identify putative regulatory motifs conserved in a set of sequences. TRES could be useful to identify evolutionarily conserved motifs in orthologous promoter sequences. Recent developments in DNA-microarray technology now allow tracking of expression of each and every gene at various snapshots during development or in response to specific stimulus. From such analysis it is possible to identify a large number of genes that show coordinate patterns of expression and comparative promoter sequence analysis of such genes can unravel common regulatory modules directing their expression. Further improvements in TRES will be necessary to make the TRES program more robust to analyze promoter sequences of a large number of genes typically identified from genome scale expression studies.

In summary, in my thesis, I have made an attempt to show how applications of simple programming designs are useful to generate new biological information. With the exponential growth of sequence and structure data, Bioinformatics will continue to play an important role in new biological discovery and in formulating intelligent questions for designing experiments. Finally, as Jacquous Monod said, "The ultimate rationale behind all purposeful structures and behaviors of living things is embodied in the sequence. And in a real sense, it is at this level of organization that the secret of life (if there is one) is to be found."



Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252:1651-1656.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science, 287:2185-2195.

Alba, M. M., M. F. Santibanez-Koref, and J. M. Hancock (1999) Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J. Mol. Evol., 49:789-797.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990) Basic local alignment search tool. J. Mol. Biol., 215:403-410.

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

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature, 290:457-465.

Andersson, L. and M. W. Freeman (1998) Functional changes in scavenger receptor binding conformation are induced by charge mutants spanning the entire collagen domain. J. Biol. Chem., 273:19592-19601.

Andrade, M. A., C. P. Ponting, T. J. Gibson, and P. Bork (2000) Homology based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol., 298:521-537.

Arguello-Astorga, G. R. and L. R. Herrera-Estrella (1996) Ancestral multipartite units in light-responsive plant promoters have structural features correlating with specific phototransduction pathways. Plant Physiol., 112:1151-1166.

Ashley, C. T., and S. T. Warren (1995) Trinucleotide repeat expansion and human disease. Annu. Rev. Genet., 29:703-728.

Avise, J. C. (1994) *Molecular markers, natural history and evolution*. Chapman and Hall, New York. pp 92-138.

Bachtrog, D., M. Agis, M. Imhof, and C. Schlotterer (2000) Microsatellite variability differs between dinucleotide repeat motifs- Evidence from *Drosophila melanogaster*. Mol. Biol. Evol., 17:1277-1285.

Bachtrog, D., S. Weiss, B. Zangerl, G. Brem, and C. Schlotterer (1999) Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol., 16:602-610.

Bairoch, A. and R. Apweiler (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res., 27:49-54.

Bateman, A., A. G. Murzin and S. A. Teichmann (1998) Structure and distribution of pentapeptide repeats in bacteria. Protein Sci., 7:1477-1480.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer (2000) The Pfam protein families database. Nucleic Acids Res., 28:263-266.

Baum, J. and B. Brodsky (1999) Folding of peptide models of collagen and misfolding in disease. Curr. Opin. Struct. Biol., 9:122-128.

Beckman, J. S. and J. L. Weber (1992) Survey of human and rat microsatellites. Genomics, 12:627-631.

Beckmann, J. S. and M. Soller (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Bio/technology, 8:930-932.

Blackwell, T. K., J. Huang, A. Ma, L. Kretzner, F. W. Alt, R. N. Eisenman, and H. Weintraub (1993) Binding of Myc proteins to canonical and noncanonical DNA sequences. Mol. Cell. Biol., 13:5216-5224.

Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science, 277:1453-1474.

Bork, P. and T. J. Gibson (1996) Applying motif and profile searches. Methods Enzymol., 266:162-184.

Brodsky, B. (1990) Fibrous proteins: folding and higher order structure. In: Gierasch, L. M. and J. King, eds. *Protein Folding*. American Association for the Advancement of Science, Washington, DC. pp 55-62.

Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol., 212:563-578.

Bucher, P. (1999) Regulatory elements and expression profiles. Curr. Opin. Struct. Biol., 9:400-407.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science, 273:1058-1073.

Buratowski, S. (1994) The basics of basal transcription by RNA polymerase II. Cell, 77: 1-3.

Burge, C. and S. Karlin (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268:78-94.

Burge, C. B. and S. Karlin (1998) Finding the genes in genomic DNA. Curr. Opin. Struct. Biol., 8:346-354.

Buscaglia, C. A., J. Alfonso, O. Campetella, and A. C. Frasch (1999) Tandem amino acid repeats from *Trypanosoma cruzi* shed antigens increase the half-life of proteins in blood. Blood, 15:2025-2032.

Carey, M. (1998) The enhanceosome and transcriptional synergy. Cell, 92:5-8.

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

Chen, Q. K., G. Z. Hertz, and G. D. Stormo (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput. Appl. Biosci., 11:563-566.

Chou K. C. (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol., 223:509-517.

Claros, M. G., S. Brunak and G. Heijne (1997) Prediction of N-terminal protein sorting signals. Curr. Opin. Struct. Biol., 7:394-398.

Claverie, J. M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. Human Mol. Genet., 6:1735-1744.

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

Corden, J. L. and M. Patturajan (1997) A CTD function linking transcription to splicing. Trends Biochem. Sci., 22:413-416.

Coward, E. and F. Drablos (1998) Detecting periodic patterns in biological sequences. Bioinformatics, 14: 498-507.

Cvekl, A., C. M. Sax, E. H. Bresnick, and J. Piatigorsky (1994) A complex array of positive and negative elements regulates the chicken alpha A-crystallin gene: involvement of Pax-6, USF, CREB and/or CREM, and AP-1 proteins. Mol. Cell Biol., 14:7363-7376.

Doolittle, R. F. (1989) Redundancies in protein sequences. In: Fasman G. D., ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 599-623.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt, J. E. Collins, R. Bruskiewich, D. M. Beare, M. Clamp, L. J. Smink, et al. (1999) The DNA sequence of human chromosome 22. Nature, 402:489-495.

Duret, L. and P. Bucher (1997) Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol., 7:399-406.

Ellegren, H. (2000a) Heterogeneous mutation processes in human microsatellite DNA sequences. Nature Genet., 24:400-402.

Ellegren, H. (2000b) Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet., 16:551-558.

Fickett, J. W. (1996) Coordinate positioning of MEF2 and myogenin binding sites. Gene, 172:GC19-GC32.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269:496-512.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science, 270:397-403.

Frech, K., J. Danescu-Mayer, and T. Werner (1997a) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J. Mol. Biol., 270:674-687.

Frech, K., K. Quandt, and T. Werner (1997b) Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem. Sci., 22:103-104.

Frech, K., K. Quandt, and T. Werner (1997c) Software for the analysis of DNA sequence elements of transcription. Comput. Appl. Biosci., 13:89-97.

Frech, K., P. Dietze, and T. Werner (1997d) ConsInspector 3.0: new library and enhanced functionality. Comput. Appl. Biosci., 13:109-110.

Freier, S. M., R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. USA, 83:9373-9377.

Freudenreich, C. H., J. B. Stavenhagen, and V. A. Zakian (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol. Cell. Biol., 17:2090-2098.

Gacy, A. M., G. Goellner, N. Juranic, S. Macura, and C. T. McMurray (1995) Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell, 81:533-540.

Garnier, J. and B. Robson (1989) The GOR method for predicting secondary structures in proteins. In: Fasman G. D., ed. *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York. pp 417-466.

Gautheret, D., F. Major, and R. Cedergren (1990) Computer modeling and display of RNA secondary and tertiary structures. Methods Enzymol., 183:318-330.

Gerber, H. P., K. Seipel, O. Georgiev, M. Hofferer, M. Hug, S. Rusconi, and W. Schaffner (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science, 263:808-811.

Ghosh, D. (2000) Object oriented Transcription Factor Database (ooTFD). Nucleic Acids Res., 28:308-310.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, et al. (1996) Life with 6000 genes. Science, 274:546-567.

Golding, G. B. (1999) Simple sequence is abundant in eukaryotic proteins. Protein Sci., 8:1358-1361.

Gorodkin, J., L. J. Heyer, and G. D. Stormo (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res., 25:3724-3732.

Graveley, B. R. and T. Maniatis (1998) Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. Mol. Cell, 1:765-771.

Green, H. and N. Wang (1994) Codon reiteration and the evolution of proteins. Proc. Natl. Acad. Sci. USA, 91:4298-4302.

Groves, M. R. and D. Barford (1999) Topological characteristics of helical repeat proteins. Curr. Opin. Struct. Biol., 9:383-389.

Gumucio, D. L., D. A. Shelton, W. Zhu, D. Millinoff, T. Gray, J. H. Bock, J. L. Slightom, and M. Goodman (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. Mol. Phylogenet. Evol., 5:18-32.

Hamada, H., M. G. Petrino, and T. Kakunaga (1982) A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA, 79:6465-6469.

Hancock, J. M. (1995) The contribution of slippage-like processes to genome evolution. J. Mol. Evol., 41:1038-1047.

Hardison, R. C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet., 16:369-372.

Harmer, S. L., J. B. Hogenesch, M. Straume, H. S. Chang, B. Han, T. Zhu, X. Wang, J. A. Kreps, and S. A. Kay (2000) Orchestrated transcription of key pathways in arabidopsis by the circadian clock. Science, 290:2110-2113.

Harr, B. and C. Schlotterer (2000) Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide under-representation. Genetics, 155:1213-1220.

Harr, B., B. Zangerl, and C. Schlotterer (2000) Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. Mol. Biol. Evol., 17:1001-1009.

Hattori, M., A. Fujiyama, T. D. Taylor, H. Watanabe, T. Yada, H. S. Park, A. Toyoda, K. Ishii, Y. Totoki, D. K. Choi, et al. (2000) The DNA sequence of human chromosome 21. Nature, 405:311-319.

Heidelberg, J. F., J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. A. Umayam, et al. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature, 406, 477-483.

Heinemeyer, T., X. Chen, H. Karas, A. E. Kel., O. V. Kel, I. Liebich, T. Meinhardt, I. Reuter, F. Schacherer, and E. Wingender (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res., 27:318-322.

Henikoff, J. G., E. A. Greene, S. Pietrokovski, and S. Henikoff (2000) Increased coverage of protein families with the Blocks Database servers. Nucleic Acids Res., 28: 228-230.

Heringa, J. and P. Argos (1993) A method to recognize distant repeats in protein sequences. Proteins, 17:391-411.

Heringa, J. and W. R. Taylor (1997) Three-dimensional domain duplication, swapping and stealing. Curr. Opin. Struct. Biol., 7:416-421.

Higo, K., Y. Ugawa, M. Iwamoto, and T. Korenaga (1999) Plant cis-acting regulatory DNA elements (PLACE) database:1999. Nucleic Acids Res., 27:297-300.

Hofacker, I. L., M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res., 26:3825-3836.

Hofmann, K., P. Bucher, L. Falquet and A. Bairoch (1999) The PROSITE database, its status in 1999. Nucleic Acids Res., 27:215-219.

Holder, A. A. (1994) Proteins on the surface of the malaria parasite and cell invasion. Parasitology, 108(Suppl):S5-18.

Hoppe, H. J. and K. B. Reid (1994) Collectins- soluble proteins containing collagenous regions and lectin domains- and their roles in innate immunity. Protein Sci., 3:1143-1158.

Huntley, M. and G. B. Golding (2000) Evolution of simple sequence in proteins. J. Mol. Evol., 51:131-140.

Ilagan, J. G., A. Cvekl, M. Kantorow, J. Piatigorsky, and C. M. Sax (1999) Regulation of alpha A-crystallin gene expression: Lens specificity achieved through the differential placement of similar transcriptional control elements in mouse and chicken. J. Biol. Chem., 274:19973-19978.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature, 409:860-921.

Jaeger, J. A., D. H. Turner, and M. Zuker (1990) Predicting optimal and suboptimal secondary structure for RNA. Methods Enzymol., 183:281-306.

Julien, J. P. and W. E. Mushynski (1998) Neurofilaments in health and disease. Prog. Nucleic Acid Res. Mol. Biol., 61:1-23

Jurka, J. and C. Pethiyagoda (1995) Simple repetitive DNA sequences from primates: Compilation and analysis. J. Mol. Evol., 40:120-126.

Kadonaga, J. T. (1998) Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines. Cell, 92:307-313.

Kajava, A. V. and S. E. Lindow (1993) A model of the three-dimensional structure of ice nucleation proteins. J. Mol. Biol., 232:709-717.

Kang, S., A. Jaworski, K. Ohshima, and R. D. Wells (1995) Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. Nat. Genet., 10:213-218.

Kantorow, M., A. Cvekl, C. M. Sax, and J. Piatigorsky (1993) Protein-DNA interactions of the mouse alpha A-crystallin control regions: Differences between expressing and non-expressing cells. J. Mol. Biol., 230:425-435.

Kashi, Y., D. King and M. Soller (1997) Simple sequence repeats as a source of quantitative genetic variation. Trends Genet., 13:74-78.

Kel, O. V., A. G. Romaschenko, A. E. Kel, E. Wingender, and N. A. Kolchanov (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucleic Acids Res., 23:4097-4103.

Kishore, U. and K. B. Reid (1999) Modular organization of proteins containing C1q-like globular domain. Immunopharmacology, 42:15-21.

Kobe, B. and A. V. Kajava (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem. Sci., 25:509-515.

Kokoska, R. J., L. Stefanovic, H. T. Tran, M. A. Resnick, D. A. Gordenin, and T. D. Petes (1998) Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (*rad27*) and DNA polymerase delta (*pol3-t*). Mol. Cell. Biol., 18:2779-2788.

Krejci, E., S. Thomine, N. Boschetti, C. Legay, J. Sketelj, and J. Massoulie (1997) The mammalian gene of acetylcholine-esterase associated collagen. J. Biol. Chem., 272:22840-22847.

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

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol. Biol. Evol., 17:1210-1219.

Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature, 390:249-256.

Lamb, P. and S. L. McKnight (1991) Diversity and specificity in transcriptional regulation: the benefits of heterotypic dimerization. Trends Biochem. Sci., 16:417-422.

Landschulz, W. H., P. F. Johnson and S. L. McKnight (1988) The Leucine Zipper: a hypothetical structure common to a new class of DNA binding proteins. Science, 240:1759-1764.

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

Lin, X., S. Kaul, S. Rounsley, T. P. Shea, M. I. Benito, C. D. Town, C. Y. Fujii, T. Mason, C. L. Bowman, M. Barnstead, et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature, 402:761-768.

Lipman, D.J. and W. R. Pearson (1985) Rapid and sensitive protein similarity searches. Science, 227:1435-1441.

Majewski, J. and J. Ott (2000) GT repeats are associated with recombination on human chromosome 22. Genome Res., 10:1108-1114.

Maniatis, T., S. Goodbourn, and J. A. Fischer (1987) Regulation of inducible and tissue-specific gene expression. Science, 236:1237-1245.

Marcotte, E. M., M. Pellegrini, T. O. Yeates, and D. Eisenberg (1999) A census of protein repeats. J. Mol. Biol., 293:151-160.

Martin, G. B., S. H. Brommonschenkel, J. Chunwongse, A. Frary, M. W. Ganal, R. Spivey, T. Wu, E. D. Earle, and S. D. Tanksley (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science, 262:1432-1436.

Maxam, A. M. and W. Gilbert (1977) A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA, 74:560-564.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. Dusterhoft, W. Stiekema, K. D. Entian, N. Terryn, et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature, 402:769-777.

McLachlan, A. D (1993) Multichannel Fourier analysis of patterns in protein sequences. J. Phys. Chem., 97: 3000-3006.

Meshi, T. and M. Iwabuchi (1995) Plant transcription factors. Plant Cell Physiol., 36:1405-1420.

Miner, J. N. and K. R. Yamamoto (1991) Regulatory cross-talk at composite response elements. Trends Biochem. Sci., 16:423-426.

Miret, J. J., L. Pessoa-Brandao, and R. S. Lahue, (1997) Instability of CAG and CTG trinucleotide repeats in *Saccharomyces cerevisiae*. Mol. Cell. Biol., 17:3382-3387.

Mironov A. A., E. V. Koonin, M. A. Roytberg, and M. S. Gelfand (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucleic Acids Res., 27:2981-2989.

Mitas, M. (1997) Trinucleotide repeats associated with human disease. Nucleic Acids Res., 25:2245-2253.

Mitas, M., A. Yu, J. Dill, T. J. Kamp, E. J. Chambers, and I. S. Haworth (1995) Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$. Nucleic Acids Res., 23:1050-1059.

Moore, H., P. W. Greenwell, C. P. Liu, N. Arnheim, and T. D. Petes (1999) Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA, 96:1504-1509.

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

Muragaki, Y., S. Mundlos, J. Upton, and B. R. Olsen (1996) Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science, 272:548-551.

Needleman, S. B. and C. D. Wunsch (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48:443-453.

Panaud, O., X. Chen, and S. R. McCouch (1995) Frequency of microsatellite sequences in rice (*Oryza sativa* L.). Genome, 38:1170-1176.

Pardue, M. L., K. Lowenhaupt, A. Rich, and A. Nordheim (1987) $(dC\text{-}dA)_n.(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. EMBO J., 6:1781-1789.

Paulson, H. L. (1999) Protein fate in neurodegenerative proteinopathies: Polyglutamine diseases join the (mis) fold. Am. J. Hum. Genet., 64:339-345.

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

Pearson, W. R. and D. J. Lipman (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85:2444-2448.

Pellegrini, M., E. M. Marcotte, and T. O. Yeates (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. Proteins, 35:440-446.

Perutz, M. F. (1999) Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem. Sci., 24:58-63.

Perutz, M. F., T. Johnson, M. Suzuki, and J. T. Finch (1994) Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. Proc. Natl. Acad. Sci. USA, 91:5355-5358.

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

Prestridge, D. S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. Comput. Appl. Biosci., 12:157-160.

Prevelige, P. and G. D. Fasman (1989) Chou-Fasman prediction of the secondary structure of proteins: The Chou-Fasman-Prevelige algorithm. In: Fasman G.D., ed. *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York. pp 391-416.

Ptashne, M. (1988) How eukaryotic transcriptional activators work. Nature, 335:683-689.

Puglisi, E. V. and J. D. Puglisi (1997) RNA structure. In: Harford J. B., and D. R. Morris, Eds, *mRNA Metabolism and Post-Transcriptional Gene Regulation*. Wiley-Liss Inc, pp 1-21.

Quandt, K., K. Frech, H. Karas, E. Wingender, and T. Werner (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res., 23:4878-4884.

Raetz, C. R. H. and S. L. Roderick (1995) A left-handed parallel β helix in the structure of UDP-N-acetylglucosamine acetyltransferase. Science, 270:997-1000.

Reese, M. G., D. Kulp, H. Tammana, and D. Haussler (2000) Genie - gene finding in *Drosophila melanogaster*. Genome Res., 10:529-538.

Richard, G. F., and B. Dujon (1996) Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae*. Gene, 174:165-174.

Richardson, J. S. and D. C. Richardson (1989) Principles and patterns of protein conformation. In: Fasman G. D., ed. *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York. pp 1-98.

Rost B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol., 266:525-539.

Saitou, N. (1996) Reconstruction of gene trees from sequence data. Methods Enzymol., 266:427-449.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen (1982) Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol., 162:729-773.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. Nature, 265:687-695.

Sanger, F., S. Nicklen, and A. R. Coulson (1977b) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA, 74:5463-5467.

Sawyer, L. A., J. M. Hennessy, A. A. Piexoto, E. Rosato, H. Parkinson, R. Costa, and C. P. Kyriacou (1997) Natural variation in Drosophila clock gene and temperature compensation. Science, 278:2117-2120.

Schorderet D. F. and S. M. Gartler (1992) Analysis of CpG suppression in methylated and nonmethylated species. Proc. Natl. Acad. Sci. USA 89:957-961.

Schug, M. D., C. M. Hutter, K. A. Wetterstrand, M. S. Gaudette, T. F. Mackay, and C.F. Aquadro (1998) The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol., 15:1751-1760.

Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res., 28:231-234.

Searls, D. B. (1998) Grand challenges in computational biology. In. Salzberg S. L., D. B. Searls and S. Kasif, eds, *Computational methods in molecular biology*. Elsevier Science B. V., pp 3-10.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchi-Shinozaki, et al. (1986) The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. EMBO J., 5:2043-2049.

Shore, P. and A. D. Sharrocks (1995) The MADS-box family of transcription factors. Eur. J. Biochem., 229:1-13.

Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell, and T. D. Petes (1997) Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. Mol. Cell. Biol., 17:2851-2858.

Sicheri, F. and D. S. Yang (1995) Ice-binding structure and mechanism of an antifreeze protein from winter flounder. Nature, 375: 427-431.

Sinden, R. R. (1999) Biological implications of the DNA structures associated with disease-causing triplet repeats. Am. J. Hum. Genet., 64:346-353.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. (1986) Fluorescence detection in automated DNA sequence analysis. Nature, 321:674-679.

Smith, T. F. and M. S. Waterman (1981) Identification of common molecular subsequences. J. Mol. Biol., 147:195-197.

Stormo, G. D. and D. S. Fields (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem. Sci., 23:109-113.

Tautz D., M. Trick, and G. A. Dover (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature, 322:652-656.

Tautz, D., and M. Renz (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12:4127-4138.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408:796-815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science, 282:2012-2018.

Thompson, J. D., D. G. Higgins and T. J. Gibson, (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.

Tinoco, I., O. C. Uhlenbeck, and M. D. Levine (1971) Estimation of secondary structure in ribonucleic acids. Nature, 230:362-367.

Tjian, R. and T. Maniatis (1994) Transcriptional activation: A complex puzzle with few easy pieces. Cell, 77:5-8.

Toth, G., Z. Gaspari, and J. Jurka (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res., 10:967-981.

Tsoka, S. and C. A. Ouzounis (2000) Recent developments and future directions in computational genomics. FEBS Lett., 480:42-48.

Uberbacher, E. C., Y. Xu, and R. J. Mural (1996) Discovering and understanding genes in human DNA sequence using GRAIL. Methods Enzymol., 266:259-281.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. (2001) The sequence of the human genome. Science, 291:1304-1351.

Wang, Z., J. L. Weber, G. Zhong, and S. D. Tanksley (1994) Survey of plant short tandem DNA repeats. Theor. Appl. Genet., 88:1-6.

Wasserman, W. W. and J. W. Fickett (1998) Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol., 278:167-181.

Wierdl, M., M. Dominska, and T. D. Petes (1997) Microsatellite instability in yeast: Dependence on the length of the microsatellite. Genetics, 146:769-779.

Wilbur, W. J. and D. J. Lipman (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA, 80:726-730.

Wolfertstetter, F., K. Frech, G. Herrmann, and T. Werner (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. Comput. Appl. Biosci., 12:71-80.

Xu, X., M. Peng, Z. Fang, and X. Xu (2000) The direction of microsatellite mutations is dependent upon allele length. Nature Genet., 24:396-399.

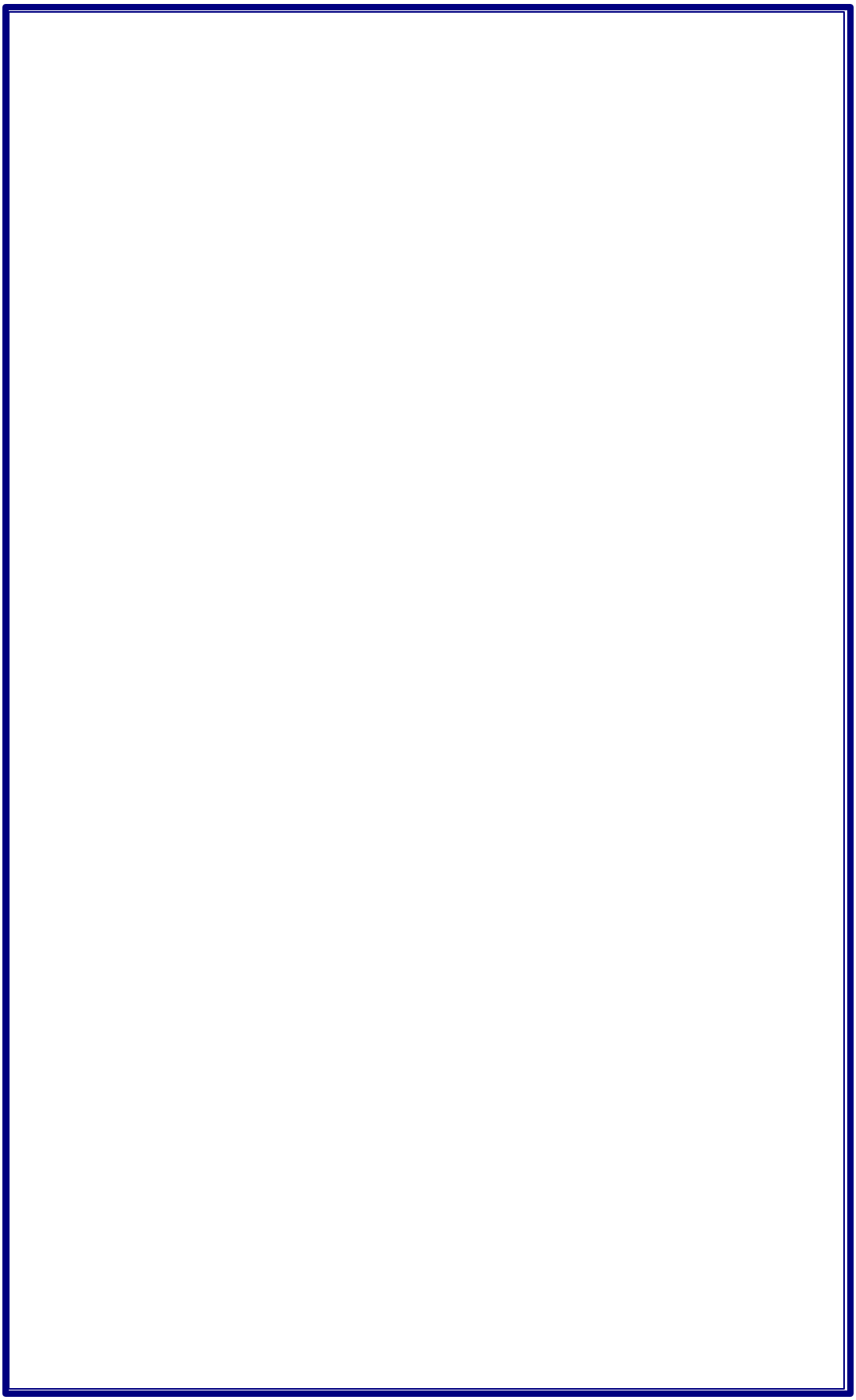
Yanofsky, M. F., H. Ma, J. L. Bowman, G. N. Drews, K. A. Feldmann, and E. M. Meyerowitz (1990) The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature, 346:35-39.

Yuh, C., H. Bolouri, and E. H. Davidson (1998) Genomic cis-regulatory logic: Experimental and computational analysis of sea urchin gene. Science, 279:1896-1902.

Zhang, Z., A. A. Schaffer, W. Miller, T. L. Madden, D. J. Lipman, E. V. Koonin and S. F. Altschul (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res., 26:3986-3990.

Zuker, M. (2000) Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol., 10:303-310.

Zuker, M., and P. Stiegler (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9:133-148.



BIO-DATA

NAME **MUKUND VYANKATESH KATTI**

PERMANENT ADDRESS c/o V. N. Katti
AT/PO Ugar Khurd, PIN 591 316
Dist. Belgaum, Karnataka, INDIA

Email: mvkatti@yahoo.com

EDUCATION:

Degree	University	Year of passing	Class
B. Sc.	Mahatma Phule Agricultural University, Rahuri, Maharashtra	1991	First Class
M. Sc. (Genetics and Plant Breeding)	Mahatma Phule Agricultural University, Rahuri, Maharashtra	1994	First Class with Distinction
Ph. D.* (Biotechnology)	University of Pune, Pune, Maharashtra	2001	Submitted the thesis

*Submitted the thesis in April 2001

FELLOWSHIPS AND AWARDS:

- Indian Council of Agricultural Research (ICAR) Junior Research Fellowship during M. Sc. studies (1991-1993).
- Qualified in National Eligibility Test (NET) conducted by Agricultural Scientists Recruitment Board, New Delhi in the subject Genetics (1994).
- Qualified in National Eligibility Test (NET) conducted by CSIR-UGC in the subject Life Sciences and was awarded CSIR Research Fellowship (1995-2000)
- NCL Research Foundation "Best Research Paper Award" for the year 2000 in the subject Biological Sciences.

PUBLICATIONS:

- **Katti M. V.**, R. Sami-Subbu, P. K. Ranjekar and V. S. Gupta (2000)
  Amino acid repeat patterns in protein sequences: Their diversity and structural functional implications. Protein Science, 9:1203-1209.

- **Katti M. V**., M. K. Sakharkar, P. K. Ranjekar and V. S. Gupta (2000)
  TRES : for comparative promoter sequence analysis. Bioinformatics, 16:739-740.

- **Katti M. V**., P. K. Ranjekar and V. S. Gupta (2001)
  Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution (*In press*).