ANALYSIS OF SIMPLE SEQUENCE REPEATS IN AMINO ACID AND NUCLIOTIDES

POORNA DAVALI[1], B.V.RAMANA[2]

[1]Department of IT, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India
email: poorna.devali1@gmail.com
[2]Assoc.professor, Department of IT, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

Abstract— **Recently, microsatellites have gained much attention due to their suggested association with cancers, ageing and various other metabolic disorders. Microsatellites are thought to have evolved mainly through polymerase slippage with variable mutation rates. It is difficult to develop a typical molecular evolutionary model that may describe genomic dynamics of these sequence elements. Microsatellites may also be the accidental sites of action by selection forces in events like genome divergence and speciation. Substantial evidence is available to describe various life stages of microsatellite evolution. Our analysis exposed that all the microsatellites (National diabetes information clearinghouse) of the KCNJ11 does not contain any mutations and these mutations also does not fall in the functional domains of the KCNJ11 thus indicating that here there is no role of microsatellites in the mutations of KCNJ11 gene.**

Index Terms— **Bio-informatics, Microsatellites, KCNJ11, GENE Analysis.**

I. INTRODUCTION

Availability of whole genome sequences and a wealth of published literature reporting the analysis of genomic sequences facilitate studies that aim at understanding various aspects of genome organization and evolution in different forms of life. The present day eukaryotic genomes pack a bulk of non-coding DNA embedded with protein coding regions. A part of this non-coding DNA plays a regulatory role, whereas the other part simply provides structural stability to the chromosomes.

Repetitiveness of nucleotide sequences is an important feature of all genomes, however, the extent to which it occurs within genomes varies greatly. The only consensus reached so far is that the amount of repetitiveness exceeds the expected values of repetitiveness[1]. Repetitive sequences are now known to play important roles in a cell, define genome structure and drive the adaptive evolution of an organism[2–4]. These sequences are broadly classified into interspersed repeats and tandem repeats, and may constitute a significant proportion of some genomes[5–7]. Tandem repeats are broadly classified into satellites, minisatellites and microsatellites, and mainly distinguished on the basis of the length of the repeating unit. Here, we critically overview the genesis and propagation of microsatellites and how the evolutionary events involving microsatellites affect the genomic transitions leading to major changes including genetic drift and speciation over long periods.

Microsatellites are remarkably constituted of small repeating units, 1–6 bp in length. Such a unit formation is structurally simple and therefore these repeats are also called as simple sequence repeats. These sequences constitute hypervariable regions of the genome and undergo structural changes through addition or removal of repeat units or through point mutations therein[8,9]. The latter event can cause imperfections in these arrays, thus leading to the formation of perfect and imperfect microsatellites[10]. The idea whether microsatellites are evolutionary junks, or useful sequences that are repeated throughout the genome has been a topic of debate in the scientific community. Evidences are being gathered in favour of the hypothesis that the simplicity of these sequences in itself is a useful attribute of the genome3 and also that they are strategically placed in the genomes. However, a recent study by Buschiazzo and Gemmell[11] indicates that for most of the microsatellites, survival in mammalian genomes is only by chance and there are no evolutionary designs behind their conservation over long evolutionary periods.

II. LITERATURE SURVEY

Apart from genes, the human genome also consists of a large number of nucleotide repeat units of size 1-6 bp repeated tandemly called Micro satellites or Simple Sequence Repeats (SSRs) or Short Tandem Repeats (STRs) [12] Micro satellites are found in all the known genomes, spanning from prokaryotes, eukaryotes and viruses and are widely distributed both in coding and noncoding regions [13,14]

Mutations in these microsatellite regions occur at much higher rate when compared with those in the rest of the genome [15] Micro satellites are known to be highly polymorphic due to the high rate of mutations in their tracts [16] These mutations can be either in the form of increase / decrease of repeat

units or in the form of single nucleotide substitutions/deletions/insertions and other events[17]. Increase or decrease of repeat units of micro satellites in coding regions might lead to shift in reading frames thereby causing changes in protein product [18] and in non-coding regions are known to effect the gene regulation [19]. Point mutations (Substitutions and Indels) are also found to occur at a higher rate in micro satellites than elsewhere[20].

Micro satellite mutations with in or near certain genes are known to be responsible for some human neurodegenerative diseases. So, we made a brief study to check whether the mutations in KCNQ1, KCNH2 and SCN5A genes, have any relation with these microsatellites repeats and the study revealed interesting results.

Micro satellite mutations with in or near certain genes are known to be responsible for some human neurodegenerative diseases. So, we made a brief study to check whether the mutations in this KCNJ11 gene have any relation with these micro satellites repeats.

III. PROBLEM STATEMENT

The problem addressed in this study is to perform computational analysis of the genes/proteins causing Neonatal diabetes using the Multiple Sequence Alignment (MSA) and to find is there any relationship with the mutations of those genes with the microsatellites. The method employed in the research reduced the complexity of using large databases, analyze the data, and produced the results in a reliable way. There are two ways of doing the analysis: in vitro research and in silico research. The in vitro research is capable of diagnosing the disease and further to suggest a treatment. The analysis of the results obtained from a laboratory experiment can be used to suggest the drugs for the cure of the disease basing on the pathogenesis of the disease. This analysis uses the data that is produced as a part of the experiments conducted on a sample only. Normally, the analysis of the results to make decisions, "achieved data is also required for conducting a comparative study to obtain better results. This is a difficult task to locate, extract, manipulate, manage and analyze huge volumes of data manually. It is quite apparent that there may be a number of causes for the disease.

The laboratory experiments may produce a small amount of data and the health professionals, in addition may have some more data at his/her disposal. With this additional data, health professionals now may be able to analyze the data and produce results to acquire knowledge about the disease and its cure. As the health professionals analyze a meager data, the results so obtained may be inaccurate and insufficient in the process of making medical decisions. Particularly in the process of drug discovery, the amount of data to be extracted, stored, manipulated, managed and analyzed is quite large. As part of these activities, a very large amount of data is to be exchanged and shared.

The in silico research uses an experimental technique performed on computer or thru an automated model. This approach is capable of extracting, storing, manipulating, managing and analyzing very huge amounts of data and produce meaningful information, which can be used to make meaningful and accurate decisions. As a part of this, large amount of data may need to be extracted from a variety of sources and in different formats. The data from various sources is mined using an appropriate mining technique and the data be made available for processing.

In the present study the sequence data of genes/proteins causing Insulin Resistance Syndrome are extracted from various sequence databases assembled and the sequence data is aligned using multiple sequence alignment method using the web based tool ClustalW. From the analysis, the names of the genes/proteins that primarily cause Neonatal diabetes are inferred. ClustalW takes the sequence data of the genes/Proteins as input and align the sequences and produce phylogenetic tree. The branches of the tree are marked with the Gene/Protein name followed by the distance of the node from the root. The Gene/Protein causing the disease can be identified from the distances. The gene that is more responsible for causing the disease is identified by studying the distances. The gene corresponding to a node with lesser distance is more responsible for the disease. Likewise the genes corresponding to the nodes with the next higher distances are considered to be responsible for the disease in the decreasing order.

IV. PROPOSED SYSTEM

Now here we have selected a genetical disorder (Neonatal diabetes) and analyzed whether there is any concurrency of mutations and microsatellites and domains. For that by reviewing the literature we found that KCNJ 11 is the gene which was the main cause of the disease so after gathering the sequence and finding the Microsatellite regions I mapped these regions to the mutations in the gene which were collected from the HGMD and also searched for the matching the regions of functional domains. This study will give a brief idea that is there is any concurrency of the mutations, microsatellites and functional domains and if so can state that the microsatellites are the main cause of mutations and if we can find the cause of these microsatellites we can easily design a* drug that can stop the microsatellites generation and thus the disease.

V. ANALYSIS

Blurry vision : Hyperosmolar hyperglycemia nonketotic syndrome is the condition when body fluid is pulled out of tissues including lenses of eye, which affects the ability of lenses to focus resulting in blurry vision.

Irritability : It is one of the sign of high blood sugar because of the inefficient supply of glucose to brain and other body organs, which makes us feel tired and uneasy.

Infections : Certain signals from the body is given whenever there is fluctuation of blood sugar (due to suppression of immune system) by frequent infections of fungal or bacterial like skin infection or UTI (urinary tract infection).

Poor wound healing : High blood sugar resists the flourishing of WBC, (white blood cell) which are responsible for body immune system. When these cells do not function accordingly, wound healing is not at good pace.

Secondly, long standing diabetes leads to thickening of blood vessels which may affect proper circulation of blood in different body parts.

Complications: If you have diabetes, your blood sugar levels are too high. Over time, this can cause problems with other body functions, such as your kidneys, nerves, feet, and eyes. Having diabetes can also put you at a higher risk for heart disease and bone and joint disorders. Other long-term complications of diabetes include skin problems, digestive problems, sexual dysfunction, and problems with your teeth and gums. Very high or very low blood sugar levels can also lead to emergencies in people with diabetes. The cause can be an underlying infection, certain medicines, or even the medicines you take to control your diabetes. If you feel nauseated, sluggish or shaky, seek emergency care.

A. SEQUENCE ALIGNMENT

In bioinformatics, a sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages.

The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than to amino acids, the conservation of base pairing can indicate a similar functional or structural role. Sequence alignment can be used for non-biological sequences, such as those present in natural language or in financial data.

Very short or very similar sequences can be aligned by hand; however, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort. Instead, human knowledge is primarily applied in constructing algorithms to produce high-quality sequence alignments, and occasionally in adjusting the final results to reflect patterns that are difficult to represent algorithmically (especially in the case of nucleotide sequences). Computational approaches to sequence alignment generally fall into two categories: global alignments and local alignments. Calculating a global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity. A variety of computational algorithms have been applied to the sequence alignment problem, including slow but formally optimizing methods like dynamic programming and efficient heuristic or probabilistic methods designed for large-scale database search. Alignment Score chart:

CLUSTAL W (Multiple Sequence Alignments)	Pair wise alignments
Sequence 1: gi 623888881re flNP_000516.31 390 aa Sequence 2: ail199236931reflNP_112648.2 390 aa Sequence 3: gi 89886327jreijNP_001034916.1 381 aa Sequence 4: gi167544261reflNP_034732.1 390 aa Sequence 5: ail 1259919441re flNP_001075067. 388 aa Sequence 6: ail 109107153Ire flXP_001089155. 390 aa	Sequences (1:2) Aligned. Score: 96 Sequences (1:3) Aligned. Score: 71 Sequences (1:4) Aligned. Score: 95 Sequences (1:5) Aligned. Score: 95 Sequences (1:6) Aligned. Score: 98 Sequences (2:3) Aligned. Score: 71 Sequences (2:4) Aligned. Score: 99 Sequences (2:5) Aligned. Score: 95 Sequences (2:6) Aligned. Score: 96 Sequences (3:4) Aligned. Score: 71 Sequences (3:5) Aligned. Score: 71 Sequences (3:6) Aligned. Score: 70 Sequences (4:5) Aligned. Score: 96 Sequences (4:6) Aligned. Score: 97 Sequences (5:6) Aligned. Score: 96

Multiple Alignments
There are 5 groups Group 1: Sequences: 2 Group 2: Sequences: 2 Group 3: Sequences: 4 Group 4: Sequences: 5 Group 5: Sequences: 6
Alignment Score 31980

The microsatellites found in KCNJ 11 gene, based on this data we draw the graphs shown in the Fig 1 and 2

Consensus	iterations	from	to	Imperfection
CAC	3	639	648	10%
CAC	4	826	837	8%
ACCT	3	904	914	9%
GGCCAA	3	1125	1142	5%

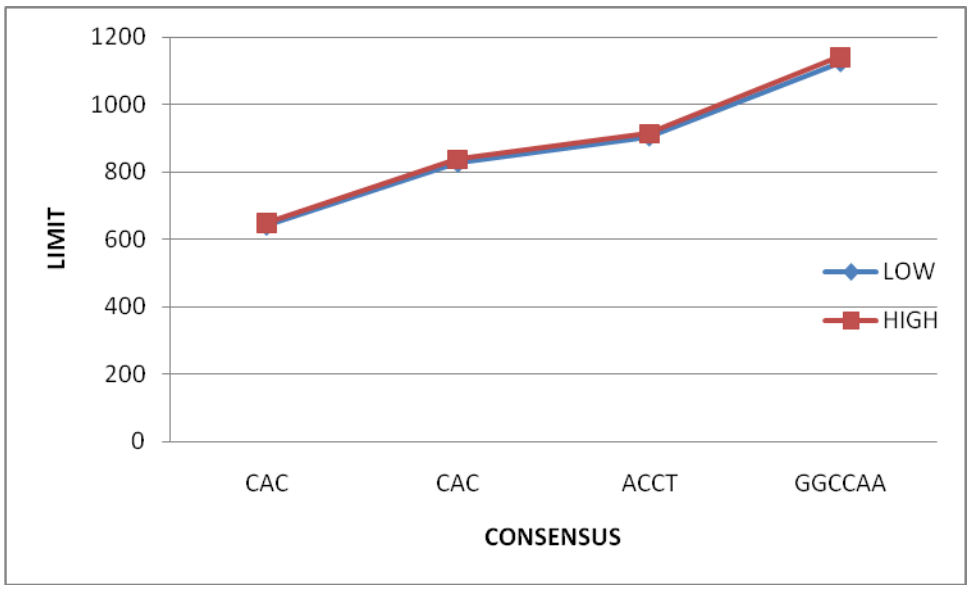

Fig 1 Consensus vs Limit

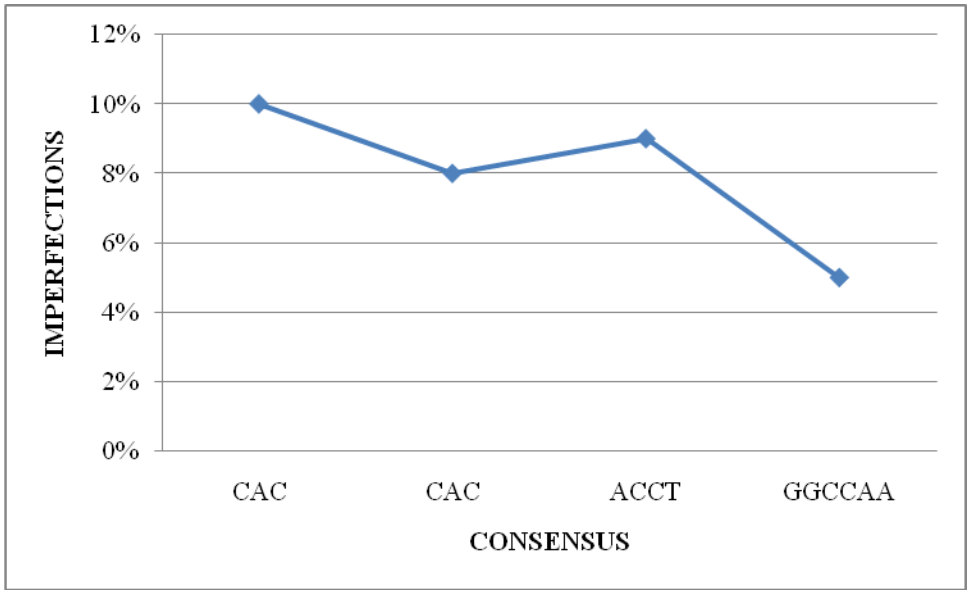

Fig 2 Consensus vs imperfections

And the mutations of KCNJ 11 gene are shown in the following table

Table 1 KCNJ11 Gene's

Accession number	Codon change	Amino acid change	Codon number
CM970815	TACG-TAA	Tyr-Term	12
CM981121	cGAG-AAG	Glu-Lys	23
CM050649	CGC-CAC	Arg-His	34
CM042726	cTTT-GTT	Phe-Val	35
CM051548	cTGC-CGC	Cys-Arg	42
CM050280	CGG-CCG	Arg-Pro	50
CM040760	CAG-CGG	Gln-Arg	52
CM050650	gGGC-AGC	Gly-Ser	53
CM050651	gGGC-CGC	Gly-Arg	53
CM040762	cGTG-ATG	Val-Mat	59
CM040761	GTG-GGG	Val-Gly	59
CMO24598	AAGt-AAC	Lys-Asn	67
CM994423	gTGG-CGG	Trp-Arg	91
CM051091	GCC-GAC	Ala-Asp	101
CMO51092	GGG-GCG	Gly-Ala	134
CMO51093	CGC-CTC	Arg-Leu	136
CM960894	CTG-CCG	Leu-Pro	147
CMO50281	AAG-AGG	Lys-Arg	170
CM050282	AAGa-AAC	Lys-Asn	170
CM050652	cATC-GTC	Arg-Cys	182
CM040763	aCGT-TGT	Arg-Cys	201
CM040764	CGT-CAT	Arg-His	201
CM043296	CCG-CTG	Pro-Leu	254
CM053288	CAT-CGT	His-Arg	259
CM051094	CCA-CTA	Pro-Leu	266
CM040765	cATC-CTC	lie-eu	296
Cmo51095	CGC-CAC	Arg-His	301

CM042727	gGAG-AAG	Glu-Lys	322
CM042728	TAC-TGC	Tyr-Cys	330
CM042729	gTTT-ATT	Phe-lle	333

VI. CONCLUSION

The mutations in the KCNJ11 are causing the neonatal diabetes mellitus. These mutations result in reduced ATP sensitivity of the KATP channels compared with the wild types. The level of channel activity defect is responsible for different clinical features: the 'mild' form confers isolated permanent neonatal diabetes whereas the severe form combines diabetes and neurological symptoms such as epilepsy, developmental delay, muscle weakness and mild dysmorphic features. So to check whether is there any relationship is there between the microsatellites and the mutations. We analyzed and found that there are no mutations in the microsatellite regions and therefore can say that the microsatellites are not responsible for mutations in the KCNJ11 gene.

References

[1]. Haubold, B. and Wiehe, T., How repetitive are genomes? BMC Bioinformatics, 2006, 7, 541.
[2]. Shapiro, J. A. and von Sternberg, R., Why repetitive DNA is essential to the genome function? Biol. Rev., 2005, 80, 227–250.
[3]. Kashi, Y. and King, D. J., Simple sequence repeats as advantageous mutators in evolution. Trends Genet., 2006, 22, 253–259.
[4]. Vinces, M. D. et al., Unstable tandem repeats in promoters confer transcriptional evolvability. Science, 2009, 324, 1213–1216.
[5]. Sharma, P. C., Grover, A. and Kahl, G., Mining microsatellites in eukaryotic genomes. Trends Biotechnol., 2007, 25, 490–498.
[6]. Piegu, B. et al., Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res., 2006, 16,1262–1269.
[7]. Warren, W. C. et al., Genome analysis of the platypus reveals unique signatures of evolution. Nature, 2008, 453, 175–183.
[8]. Eckert, K. A. and Hile, S. E., Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol. Carcinog., 2009, 48,379–388.
[9]. Shah, S. N., Hile, S. E. and Eckert, K. A., Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res., 2010, 70, 431–435.
[10]. Kofler, R., Schlotterer, C., Luschutzky, E. and Lelley, T., Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics, 2008, 9, 612.
[11]. Buschiazzo, E. and Gemmell, N. E., Conservation of human microsatellites across 450 million years of evolution. Genome Biol. Evol., 2010, 2, 153–165.
[12]. Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109: 365-371
[13]. Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967-981.
[14]. Sreenu VB, Kumar P, Nagarajaram HA (2007) Simple sequence repeats in mycobacterial genomes. J Biosci 32: 3-15.
[15]. Ellegren, H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet 24:400-402.
[16]. Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11:424-429.
[17]. Fan H, Chu JY (2007) A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5: 7-14.
[18]. Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21: 991-1007.
[19]. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER (2005) Microsatellite instability regulates transcription factor binding and gene expression. PNAS 102: 3800-3804.
[20]. Sibly RM, Meade A, Boxall N, Wilkinson MJ, Corne DW, etal. (2003) The structure of interrupted human AC microsatellites. Mol Biol Evol 20: 453-9.
[21]. Mudunuri SB, Nagarajaram HA (2007) IMEx:Imperfect Microsatellite Extractor. Bioinformatics 23: 1181-1187.
[22]. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, etal. (2004) SMART 4.0: Towards genomic data integration. Nucleic Acids Res 32: D142-4.