Sayak Ganguli et al /J. Pharm. Sci. & Res. Vol.4(1), 2012, 1692-1696

ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

# Using SVM for Identifying Epigenetic Patterns in Microsatellites in Human Sex Determining Genes and its Homologues

**Sayak Ganguli, Sasti Gopal Das, and Abhijit Datta**
*DBT-Centre for Bioinformatics, Presidency University, Kolkata*

**Abstract**
Microsatellite or Simple sequence repeats (SSRs) are found in most organisms, and occupy about 3% of the human genome. They are 1-6bp long repeated sequences which occur in several forms such as mono, di, tri, tetra, penta, hexa nucleotide repeats and are important in genome organization and function. An epigenetic change (DNA methylation) in microsatellite region has been reported to cause several diseases (especially cancer) by transcriptional silencing. This work is concerned with systematic epigenetic analysis of microsatellite region in sex-determining genes (Upstream & coding sequence) of human & other related species. Several microsatellite repeats were analyzed using SVM based Tools. It was observed that mono (A/T/G/C)n, di (TG/AC)n, tri (GCG/AGC)n are more frequent than tetra & penta nucleotide repeats.
The result of CpG region detection showed that few genes have no CpG islands or islets in their microsatellites of the upstream & coding sequences, except two genes-AR & UBE1. CpG regions were classified into two types: CpG islet (<200bp) and CpG island (>=200bp). All other regions which are devoid of such sequences are considered to be CpG orphans which have GC content <40%. The association of CpG islands within and near the microsatellite regions signifies that the mechanisms of inheritance closely follow the occurrences of duplication in the genome thus rendering a higher level of control on genome function.
**Keywords**
Sex-Determining Genes, Microsatellite, Computational Epigenetics, Methylation, CpG region

## BACKGROUND

**Epigenetics** [epi (Greek)-over or above, genetics] is defined as the study of the heritable changes in gene expression that occur without a change in DNA sequence. E*pigenetics* can also be used to describe environmental hereditary changes that can possibly influence the development of an organism [1]. Many works provide evidence that epigenetic changes play a critical role in the development of certain human diseases, such as, neurodevelopmental disorders, cardiovascular diseases, type-2 diabetes, etc.
These heritable epigenetic changes include DNA Methylation, post-translational modifications of histone tails (acetylation, methylation, phosphorylation etc.) and higher order packaging of DNA around nucleosomes.
DNA methylation is the covalent addition of a methyl group to 5$^{th}$ position of cytosine, which is largely confined to CpG dinucleotides. DNA methylation is accomplished through the activity of specific enzymes, the DNA methyltransferases (DNMTs), which transfer a methyl group, in coordination with the universal methyl donor *S*-adenosylmethinine (SAM), to the cytosine of CpG dinucleotides (2).
Recent investigations suggest that butyrate, diallyl disulfide, and sulforaphane can inhibit histone deacetylase enzymes and alter the expression of specific genes (3).
It is becoming clear from several works, that microsattelite repeats are important in genomic organization and function and may be associated with diseases [4]. Epigenetic changes occur in the microsatellite region of upstream & gene sequence causes several cancer phenotypes such as breast, uterine, endometrial, colorectal & prostate cancer [5]. Some diseases related to the epigenetic modification of microsatellite repeats are Huntington's disease, Kennedy's Disease, Colon cancer and Prostate cancer [6].
The future potential of the epigenome is wide-ranging. The human epigenome project has already started to impact the various epigenetic research endeavours around the world.
Epigenetic studies of microsatellites in the coding and non coding regions of sex determining genes probably in near future might be able to shed light on the control of dosage compensation mechanisms in most organisms.

## METHODOLOGY

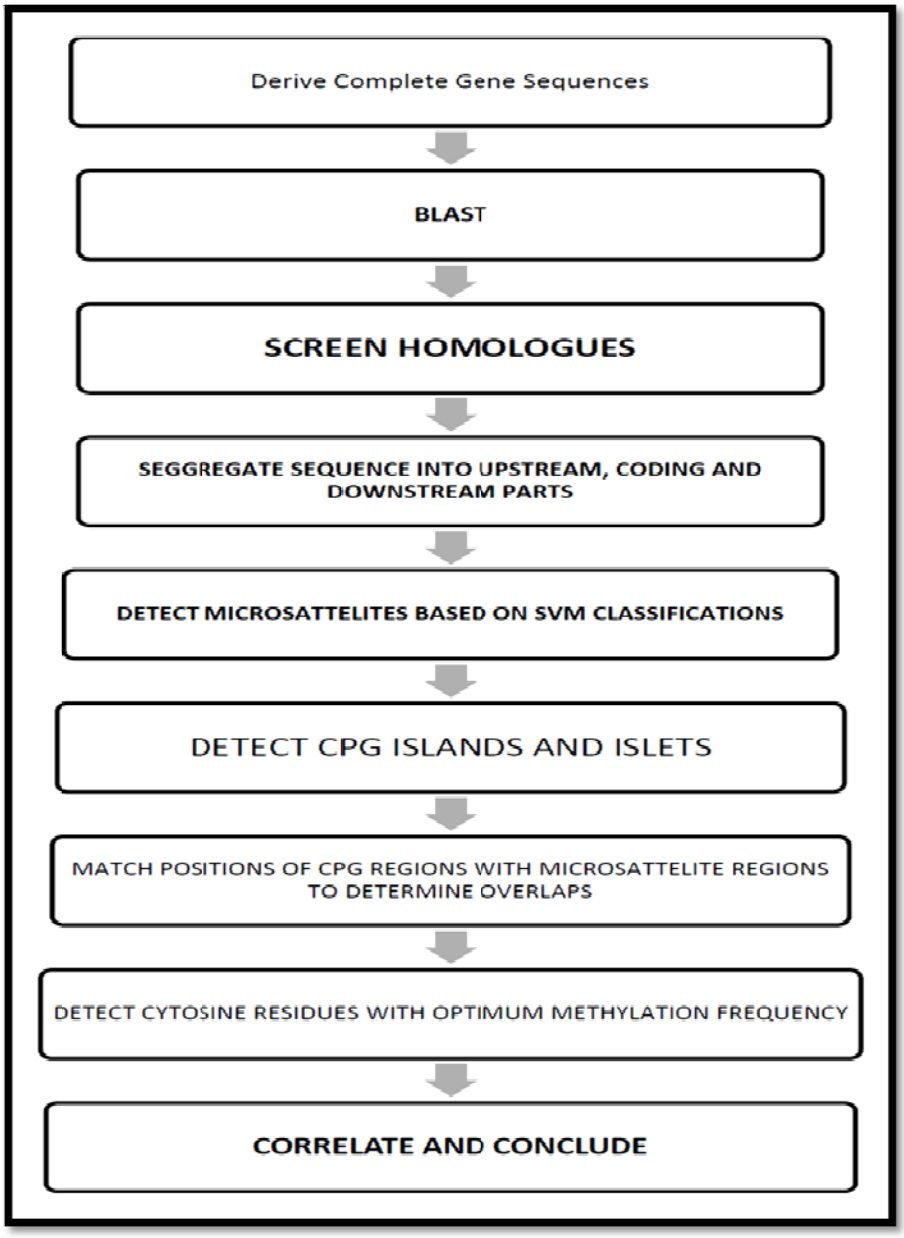


**FIG 1: Workflow for Determining Microsattelites and CPG Island in Sex Determining Genes of Human and Its Homologues.**

This miRNA::mRNA duplex is further classified with (SVMlight) for a post processing filtering. This is done by utilizing a machine-learning algorithm based on support vector machine (SVMlight) that can be used as post processing software for filtering the targets. The prediction system is trained with the experimentally supported animal miRNA targets found in TarBase [7]. Each miRNA: target interaction is mapped into a feature vector in a feature space. The feature space includes various frequencies in the interacting miRNA: target pairs. We use a feature selection procedure to filter out those features with low discriminating abilities, resulting in feature space consisting of 14 features. Vapnik first introduced (SVMs) as a class of supervised learning algorithms. Given a set of labeled training feature vectors (in our case, the positive and the negative miRNA: target interaction pairs), an SVM learns to discriminate between the two classes. The result is a trained model that can be used to classify unlabeled inputs.

## RESULTS

The analysis of the sequence provides some interesting results which are summarized below:

- Detection of multiple CpG island & CpG islet in the upstream region & coding sequence indicating a high probability of methylation.
- Detection of multiple microsatellite repeats present in the upstream & coding sequence containing methylated Cytosines.

The sequences exhibiting the above characteristics were *Homo sapiens* AR Gene, *Mus musculus* AR Gene, *Rattus norvegicus* AR Gene, Homo sapiens ZNF711 Gene, RBM10 Gene of *Equus caballus, Macaca mulatta and Mus musculus.*

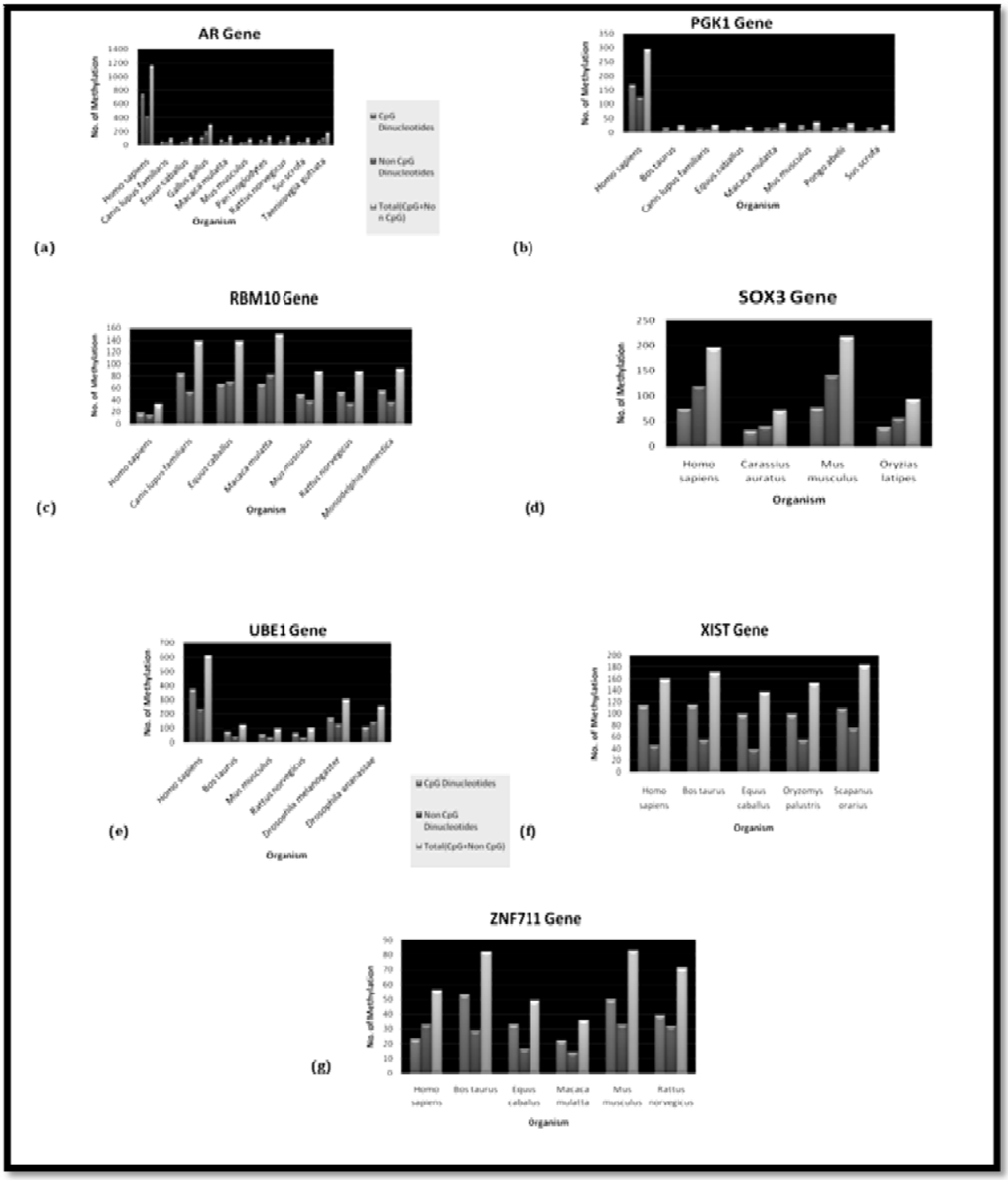


**Fig 2: (a – g) Graphical representation of the CpG & Non CpG Dinucleotides present in coding sequences of Human Sex Determining Genes and its homologues.**

**Table 1: Summary of Epigenetic Analysis of Microsatellites.**

| *Sr No* | *Gene name* | *Organism & Gi No.* | *Microsatellite Repeat & Region* | *Repeats residing in the CpG Region* | *No. of Methylated Cytosines present in the Repeats* |
|---|---|---|---|---|---|
| 1 | AR Gene | *Homo sapiens* (>ref\|NG_009014.1\|) | (GCG)17 & [7485-7535] | YES(7237-7727) | 18 |
| | | | (AC)15 & (106697-106726) | None | 1 |
| | | | (AC)17 & (156708-156741) | None | 1 |
| 2 | AR Gene | *Canis lupus familiaris* (>gi\|35384731\|) | (GCA)10 & (92-121) | None | 2 |
| | | | ( GGC)10 & (1661-1690) | YES(1438-1886) | 11 |
| 5 | AR Gene | *Mus musculus* (>gi\|118129906\|) | (CGG)5 & (149-163) | None | 5 |
| 6 | AR Gene | *Pan troglodytes* (>gi\|57113914\|) | (GGC)11 & (1290-1322) | YES(1043-1515) | 11 |
| 7 | AR Gene | *Rattus norvegicus* (>gi\|6978534\|) | (CGG)5 & (1110-1124) | YES(1017-1917) | 5 |
| 8 | AR Gene | *Taeniopygia guttata* (>gi\|115529239\|) | (CGCC)3 & (456-467) | YES(200-1017) | 3 |
| 9 | PGK1 Gene | *Homo sapiens* (>gi\|210032246\|) | (AC)6 & (21189-21200) | None | 1 |
| 10 | RBM10 Gene | *Equus caballus* (>gi\|194227850\|) | (GGC)5 & (437-451) | YES(290-641) | 4 |
| 11 | RBM10 Gene | *Macaca mulatta* (>gi\|109130670\|) | (GGC)5 & (224-238) | YES(200-431) | 4 |
| 12 | RBM10 Gene | *Mus musculus* (>gi\|21704123\|) | (GGC)5 & (344-358) | YES(331-529) | 4 |
| 13 | UBE1 Gene | *Drosophila ananassae* (>gi\|194756439\|) | (AGC)5 & (51-65) | None | 2 |
| 14 | ZNF711 Gene | *Homo sapiens* (>gi\|95147562\|) | (GGC)6 & (134-181) | None | 13 |

## DISCUSSION

The results of **Homology Search** show that the sequences share a great degree of sequence similarity. Species such as ***Mus musculus*** (House Mouse), ***Canis lupus familiaris*** (Dog), ***Equus caballus*** (Horse), ***Pan troglodytes*** (Chimpanzee), ***Macaca mulatta*** (Rhesus Monkey), ***Sus scrofa*** (Pig) and ***Bos taurus*** (Cattle) show high sequence similarity with Sex Determining Genes of ***Homo sapiens*** (Human). This is a very interesting observation since some members of the obtained groups are evolutionarily distant from the human stock.

The results of **the various SVM based Tools** show that the **CpG islands** and **CpG islets** occur in high frequency in the coding sequence. The number of CpG islets is greater than the number of CpG islands. A group of species having larger CpG island ratio are *Homo sapiens* (>ref|NG_009014.1|), *Gallus gallus* (>gi|91680855|), *Macaca mulatta* (>gi|74136372|), *Rattus norvegicus* (>gi|6978534|), *Pan troglodytes* (>gi|57113914|), *Mus musculus* (>gi|156766076|), *Drosophila melanogaster* (>gi|45 549093|), *Drosophila ananassae* (>gi|194756439|).

Some upstream sequences show CpG regions in some species such as *Homo sapiens* (>gi|95147562|) ZNF711 Gene, *Homo sapiens* (>gi|218156321|) UBE1 Gene, *Canis lupus familiaris* (>gi|35384731|) AR Gene, *Pongo abelii* (>gi|197099639|) PGK1 Gene and *Canis lupus familiaris* (>gi|74007806|) PGK1 Gene. Thus this observation leads us to believe that these genes are possibly under gene regulation through epigenetic cascades. However, upstream sequences of other genes under study do not exhibit the presence of CpG islands or islets. In case of downstream region of coding sequence no CpG region was identified in all the genes under study. This result is in conformation with

experimental results which provides information that downstream epigenetic response modules do not function in gene expression.

The **(CAG)**n repeats found in the sequence of Androgen Receptor gene of all homologous species resides in a fixed location within the coding sequence which provides evidence that Sex determination in all members under study is a highly conserved phenomenon to the extent that its regulation is specific and can be treated as a phylogenetic signature.

There are few repeats found in the upstream region of ***Homo sapiens*** AR and ZNF711 Gene, ***Drosophila melanogaster*** UBE1 Gene and ***Rattus norvegicus*** ZNF711 Gene. But the other species contain no repeats in their upstream region. Three repeats were found in the downstream region of Human Androgen receptor gene. The repeats are $(AAAAC)_2$, $(AAAT)_6$ and $(TACA)_6$. Coding sequence of most genes exhibits a high frequency of Methylated Cytosine residues in the CpG islands and islets (Figure-2). The number of methylated Cytosines is higher in the Sex determining gene of *Homo sapiens* than the other related sequences. This is a very interesting observation that tells us that a chance of Epigenetic alteration is much more frequent in this region.

Upstream region of a group of genes contain a large number of methylated Cytosines, they are- *Homo sapiens* (>ref|NG_009014.1|) ZNF711 Gene, *Pongo abelii* (>gi|1970996391|) PGK1 Gene and *Canis lupus familiaris* (>gi|74007806|) PGK1 Gene. Few genes contain methylated Cytosines in their promoter region. Therefore it suggests that there is high probability of epigenetic modification which causes gene regulation.

It was observed that a group of genes have microsatellite repeats in the CpG region and methylated Cytosines is present in the repeats. The genes residing in this group are *Homo sapiens* AR Gene, *Mus musculus* AR Gene, *Rattus norvegicus* AR Gene, Homo sapiens ZNF711 Gene, RBM10 Gene of *Equus caballus, Macaca mulatta and Mus musculus.* Upstream region of *Homo sapiens* (>gi|951475621|) ZNF711 Gene contains a repeat which resides in the CpG region and methylated Cytosines is present in this repeat.

## CONCLUSION

It can be concluded that there is a possibility of epigenetic modification in the microsatellites of sex determining genes which show characteristics of evolutionary conservedness and leads us to believe that the phenomenon of Sex determination, Dosage compensation and its related event speciation all follow the epigenetic dogma where microsatellites serve as landmarks for determining epigenome boundaries for gene regulation cascades.

## ACKNOWLEDGEMENT

The authors acknowledge the financial support provided by Department of Biotechnology, Government of India.

## REFERENCES

[1] Holliday R. DNA methylation and epigenetic defects in carcinogenesis. Mutat Res. 1987 Dec;181(2):215–217

[2] Poirier LA. Methyl group deficiency in hepatocarcinogenesis. Drug Metab Rev 1994;26:185–99.

[3] Dashwood RH, Myzak MC, Ho E. Dietary HDAC inhibitors: time to rethink weak ligands in cancer chemoprevention? Carcinogenesis 2006;27:344–9.

[4] Lander, E. S. The new genomics: Global views of biology. Science 274, 536±539 (1996)..

[5] Eder Jorge Oliveira, Juliano Gomes Pádua, Maria Imaculada Zucchi, Roland Vencovsky and Maria Lúcia Carneiro Vieira. Origin, evolution and genome distribution of microsatellites *Genetics and Molecular Biology*, 29, 2, 294-307 (2006)

[6] Moxon, E. R. and Wills, C., 1999. DNA Microsatellites: Agents of EvolutionScientific American, Jan 1999, 94-99.