EVALUATION OF MICRORNA SEQUENCE COMPOSITION AND EXPRESSION

A THESIS SUBMITTED TO
THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFULLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

By
Koray Doğan Kaya
August 2005

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Erkan Mumcuoğlu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Cengiz Yakıcıer

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Özlen Konu

Approved for the Institute of Engineering and Sciences:

_____

Prof. Dr. Mehmet B. Baray

Director of Institute of Engineering and Sciences

ABSTRACT

EVALUATION OF MICRORNA SEQUENCE COMPOSITION AND EXPRESSION

Koray Doğan Kaya

M.S. in Molecular Biology and Genetics

Supervisors: Asst. Prof. Dr. Özlen Konu, Asst. Prof. Dr. M. Cengiz Yakıcıer

August 2005, 70 pages

MicroRNAs are small, ~22nt in length, ribonucleic acids, which bind to 3'UTR regions of mRNAs by base complementation. They play crucial roles in the regulation of embryonic development, and in many cellular events such as proliferation, apoptosis, and differentiation. Previous bioinformatics studies have concentrated on the prediction and transcriptional regulation of miRNA and target sequences. In the present study, the dinucleotide and trinucleotide composition of premature and mature microRNAs were evaluated. Moreover, the significance of biased dinucleotide usage was determined using a randomization approach. The scarcity of CpG motif, a generalized characteristic of mammalian genomes, was found to be a prominent feature also in miRNA sequences. Furthermore, miRNAs that contain one or more CpG motif(s) were found to be expressed at relatively lower levels in multiple adult tissues, except in the brain, based on existing human and mouse microarray datasets. The implications of the negative correlation between the CpG usage and the average expression level of miRNAs were discussed in the light of mammalian immune defenses and methylation characteristics.

Keywords: microRNA, microarray, bioinformatics, CpG dinucleotide, randomization

ÖZET

MIKRORNA'LARIN DİZİ BİLEŞENLERİ VE İFADE DÜZEYLERİNİN ARAŞTIRILMASI

Koray Doğan Kaya

M.S. in Molecular Biology and Genetics

Tez Danışmanları: Yard. Doç. Dr. Özlen Konu, Yard. Doç. Dr. M. Cengiz Yakıcıer

Ağustos 2005, 70 sayfa

MikroRNA'lar, mesajcı RNA'ların 3' ucunda yer alan protein kodlamayan bölgelere (3'UTR) baz tamamlama ile bağlanan yaklaşık 22 nükleotid uzunluğundaki ribonükleik asitlerdir. miRNA'lar embriyonik gelişim, hücre bölünmesi ve apoptoz gibi birçok hücresel olayda önemli rol oynarlar. Daha önceki biyoenformatik çalışmalar, mikroRNA'ların ve hedef mesajcı RNA'ların tahmini ve ifade kontrolleri üzerinde yoğunlaşmıştır. Bu çalışmada ise, olgun ve öncül mikroRNA'ların ikili ve üçlü baz kompozisyonları değerlendirildi. Ek olarak, bir randomizasyon yaklaşımı kullanılarak ikili baz komposizyonundaki sapmaların istatistiksel onemi de araştırıldı. Memeli genomuna ait özgün bir karakter olan CpG motif nadirliğinin miRNA sekanslarınında belirgin bir özelliği olduğu ortaya çıkarıldı. Aynı zamanda, yayımlanmış insan ve fare mikroarray verilerine dayanarak, beyin haricindeki dokularda, bir veya birden fazla CpG motifi içeren miRNA'ların bu motifi içermeyenlerden daha az miktarda ifade edildiği de gösterildi. miRNA'ların CpG kullanımı ve ortalama ifade düzeyleri arasındaki negatif korelasyonun neden ve sonuçları memeli bağışıklık ve metilasyon mekanizmaları açılarından tartışıldı.

Anahtar Sözcükler: mikroRNA, mikrodizilim, biyoenformatik, CpG dinükleotid, randomizasyon

ACKNOWLEDGEMENTS

I wish to thank to my supervisors Dr. Özlen Konu and Dr. Cengiz Yakıcıer, for their invaluable guidance and support during my studies.

I would like to thank to Dr. Erkan Mumcuoğlu for sharing his time to read and evaluate my thesis. He helped me see whatelse can be done for the perfection of my works.

I wish to thank to Prof. Mehmet Öztürk for his endless belief in my success.

I would like to thank to my mother Selviye Kaya for her help and support. I never lived a hard time since she was always with me.

I wish to thank also to Dr. Uygar Tazebay for helping me keep my motivation in high levels by his discussions and friendship.

I would like to thank to Dr. Rengül Çetin Atalay for her crucial help.

I wish to thank my special friends Biter, Ertuğrul and Raşit for their help.

Finally, I would like to thank all Bilkent MBG family...

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BM : Bone Marrow
Br : Brain
He : Heart
K : Kidney
Li : Liver
Lu : Lung
Pa : Pancreas
Pr : Prostate
SM : Skeletal Muscle
Sp : Spleen
Th : Thymus
FC : Frontal Cortex
Ly : Lymph
Co : Colon
HL : Hela S3
Ce : Cervix
Bl : Bladder
Te : Testicle
AC : Adrenal Gland
U : Uterus
Bt : Breast
F : Fallopian
SI : Small Intestine
Pl : Placenta
Ov : Ovary

CHAPTER 1: INTRODUCTION

1.1 MicroRNAs and their discovery

MicroRNAs, which are small (21-22 nt) RNAs, play crucial roles in animals or plants during many cellular processes including differentiation, via targeting mRNAs for translational repression or cleavage (Bartel, 2004). The founding member of these recently discovered non-coding small RNAs is *lin4*. In 1993, Ambross and colleagues reported that *lin4*, a gene previously discovered as having a function in the timing of *C.elegans* larval stages, was a couple of non-coding RNAs, 61 and 22 nt in length. The question of how *lin4* regulated the developmental timing has been answered by the discovery that *lin4* had multiple complementary sites on the 3' UTR region of *lin14* mRNA (Lee et al., 1993; Wightman et al., 1993). This particular UTR region had previously been hypothesized to play an important role in the repression of *lin14* by the *lin4* product.

1.2 Tanscriptional control and maturation of microRNAs

The importance of miRNAs in regulation of the timing of various stages of development points to an intricate expressional regulation. In other words, some microRNAs should not always be expressed because they are involved in the temporal regulation of development while expression of some others should be restricted to particular tissues. Recent studies provide clues to the regulation of miRNA expression. More than half of the known mammalian microRNAs are located within the introns of protein coding and non-coding transcription units whereas ~10% of them are expressed among the exons of non-coding long transcription units (Rodriguez et al., 2004). The primary transcripts of microRNAs are known as primary microRNAs, pri-miRNA (Lee et al., 2002; Bartel, 2004). An RNase III endonuclease, Drosha, cuts the primary microRNA in both strands at the bases near the stems of precursor microRNAs (pre-microRNA), allowing them to

form stem-loop structures, 60-70 nt in length (Lee et al., 2003; Pasquinelli et al., 2005, Figure 1).

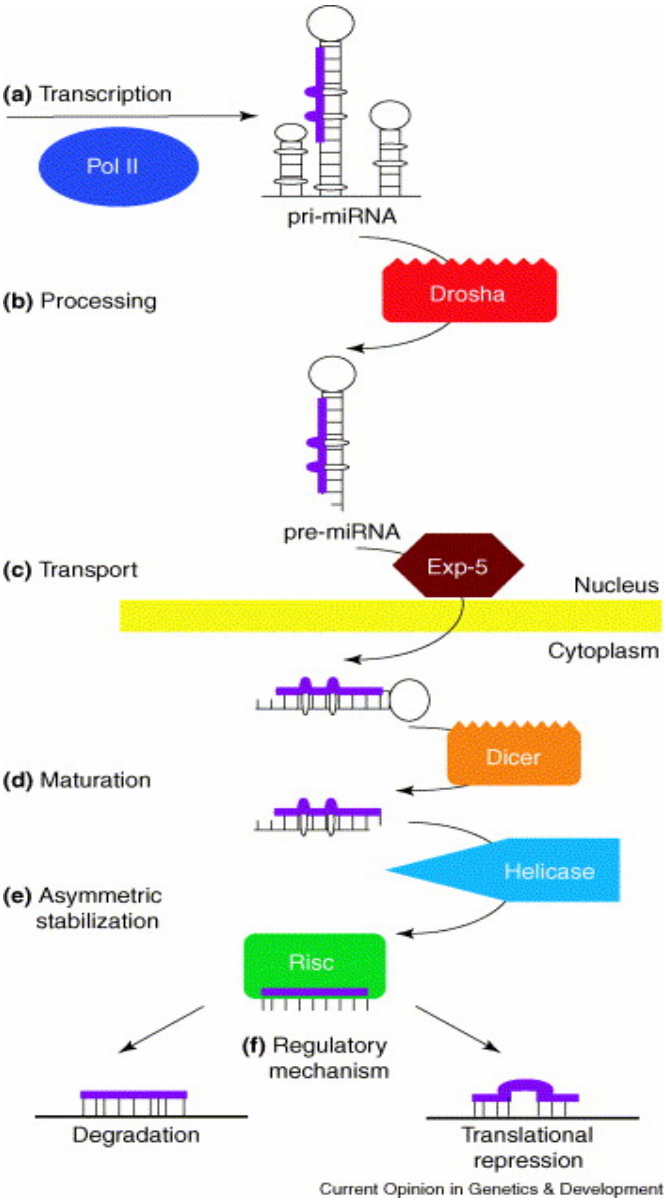

Figure 1.1 Graphical representation of the transcription and maturation of miRNAs (Pasquinelli et al., 2005).

Pre-microRNAs are actively transported to the cytoplasm by RanGTP and the export receptor exportin 5 (Yi et al., 2003; Lund et al., 2004; Bartel, 2004). After transportation to the cytoplasm from the nucleus, another RNaseIII endonuclease, Dicer, cuts the pre-microRNA producing mature microRNAs (Lee et al., 2002).

1.3 Patterns of microRNA expression

MicroRNA expression is tightly regulated, temporally and spatially. For example, *lin4* and *let7*, two stage-specific microRNAs, are expressed only in specific stages (temporally controlled expression) during the development of *C.elegans* (Pasquinelli et al., 2000). Certain microRNAs are cell-specific, such as *miR-1*, primarly found in the mammalian heart (Lagos-Quintana et al., 2002); and many other examples exist in the literature with regards to the stage- and tissue-specific expression of miRNAs (Lau et al,. 2001; Lagos-Quintana et al., 2002; Lim et al,.2003a; Chen et al,.2004).

There have been several attempts to understand the complicated expression patterns of microRNAs (Sun et al., 2004; Valoczi et al.,2004; Babak et al., 2004; Rodriguez et al., 2004; Calin et al., 2004; Liu et al., 2004; Schmittgen et al., 2004; Sempere et al., 2004; Liang et al., 2004; Baskerville et al., 2005). Among these, Valoczi et al. (2004) utilized the northern blot analysis with the complementary oligonucleotides, which contained bases at every third position substituted by the corresponding LNA[1]s (locked nucleic acid). This method due to its high sensitivity has the potential to be used for future microarray profiling of microRNAs. On the other hand, Rodriguez et al. (2004) used computational techniques only while others have made use of the microarray studies to elucidate expression properties. Although there are different approaches towards using microarrays in detection of mature miRNAs, the relative efficiency and accuracy of these approaches are yet to be compared in detail (Sun et al., 2004).

Sequence similarity searches between miRNAs indicate that their expression should be highly complicated yet correlated since similar microRNAs can act on 3' UTR targets with common motifs. Sun et al. (2004) has indicated that organ-specificity among miRNAs is a prevalent event. According to their study, *mir133a*,

[1] An LNA is an ribonucleotide analogue in which the furanose ring of the sugar-phosphate backbone is stabilized in the N conformation by the introduction of a 2'-O, 4'-C methylene bridge. This conformation favors the type A Watson-Crick duplex model in LNA-DNA and LNA-RNA base pairing (Valoczi et al., 2004)

mir133b, *mir1d* and *mir296* were expressed only in heart and skeletal muscle while *mir192*, *mir194*, *mir204*, *mir215*, and *mir216* were expressed only in kidney in humans. Neither of these was expressed in spleen, lung or prostate. The definition of a common regulatory sequence conserved across species upstream of miRNAs resulted in the discovery of a binding site for the proto-oncogene *ets1* located in the upstream region of *mir204* (1155bp). Indeed, *Est1* is a key transcription factor in normal kidney development (Razzaque et al., 2001; Cederberg et al., 1999; Gomez et al., 1999).

Babak et al. (2004) also have detected tissue-specific miRNAs in human and mouse tissues via probing with flourescent labeled RNAs from different tissues to the RNAs complementary to miRNAs on arrays. Their results were in accord with earlier studies (Kim et al., 2004; Sempere et al., 2004; Houbaviy et al., 2003; Liu et al., 2004). Particularly interesting was that the number of brain-specific miRNAs was the highest while most adult miRNAs accomplished their specialized regulatory tasks in unique tissues.

Although traditional wet-lab techniques are useful in profiling the tissue-specific expression levels of microRNAs, it is also possible to infer similar or supportive information by tracing their transcription units. Rodriguez et al. (2004) characterized the genomic positions of 232 mammalian miRNAs in terms of the transcription units they belonged to. Comparison of the computer derived data with the experimental studies suggests that tracing of the host genes instead of miRNAs themselves may be an easier and convincing way of understanding miRNA expression patterns. The search across species for conserved transcription factor binding sites will further advance our grasp of miRNA expressional regulation (Sun et al., 2004).

1.4 Genomic signature

Genomic similarities are extracted using information derived from large, non redundant genomic sequence assemblies. One method to define the genomic

similarity among a set of sequences is a profile generated based on the relative dinucleotide abundances:

$$(p^*_{xy}=f^*_{xy}/f^*_x f^*_y)^2$$

where XY is a dinucleotide motif, *p* is relative abundance of the dinucleotide; and *f* is the frequency (Gentles and Karlin, 2001). The collection of *p* values represents a genomic signature; and it is considered unique and stable for individual genomes (Karlin et al., 1997; Gentles and Karlin, 2001).

The most striking discovery made via the genomic signature comparisons has been that the TA dinucleotide motif is underrepresented in most of the genomes while the CG motif is the most underrepresented one in vertebrate genomes, hence in *H.sapiens*, but not in invertebrates (Karlin et al., 1997; Gentels and Karlin, 2001).

1.5 Relationship between GC content and expression level in mammals

Although the genomic signature has been considered relatively stable within the human genome (Karlin and Gentles, 2001), a mosaic structure of isochores, which could be defined as long DNA strings homogenous in base composition, exists in vertebrate genomes (Bernardi et al., 2002). Among different isochores, GC% rich ones have been shown as also the gene rich ones (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak et al., 1996; Lander et al., 2001).

It has been a debate whether there is a correlation between GC content and expression levels in humans (Bernardi, 1993; Gonçalves et al., 2000; Ponger et al., 2001; Galtier et al., 2001; Saccone et al., 2002). Recently, the third nucleotide position GC content and expression levels of mRNAs were shown to be correlated weakly and in a rather non-linear fashion in rodents (Konu and Li, 2002). A similar correlation was demonstrated to exist in humans as well (Arhondakis et al., 2003). Even among the GC3 richest genes, the expression of the housekeeping genes was

[2] The star indicates that the reverse complement of the sequence is concatanated to it.

shown to be significantly higher than that of tissue specific genes (Arhondakis et al., 2003). Although no apparent correlation between intron length and expression level of the genes in any frame has been reported earlier (Castillo-Davis et al., 2002), with the increasing GC3 content the intron lengths were shown to get shorter (Arhondakis et al., 2003).

Furthermore, the quantification of the average expression levels provided information towards the bias that existed in favor of more GC3 (Vinogradov, 2003). A graph of gene-specific GC content versus the number of tissues with expression suggested that tissue-specific genes have GC rich and GC poor peaks whereas the house keeping genes have only one peak skewed towards a value higher than the average GC%. On the other hand, GC% content did not vary extensively between tissue specific genes and housekeeping genes since all the peaks were in close proximity to each other on the GC content scale.

1.6 CpG motif and immunity

Individual species have unique genomic signatures thus it is possible that a foreign genomic piece is recognized by the vertebrate immune systems (Krieg, 1996; Klinman et al., 1996; Stacey et al., 1996; Yi et al., 1996).

CpG motif, a highly underrepresented one in the human genome, is a hallmark of *H. sapiens* genomic signature, as described in Section 1.4. Thus, this motif is the strongest candidate for a role in modulating immune system response in the human body. According to the recent studies, toll-like-receptors (TLRs) are responsible for the recognition of CpG containing motifs in oligo DNA or dsRNAs and for stimulation of the proinflammatory response (Dalpke 2002; Ahmad-Nejad et al., 2002).

Interestingly, Sugiyama et al. (2005) have shown that ssRNAs also could initiate immune response. The demonstrated rule was that the oligoribonucleotides with CpG motif(s) must also have contained poly(G) tail. Although ssRNAs

stimulate immune responses in special cases stated above, DNAs with multiple copies of CpG can also stimulate immune response (Zhang, 2005).

1.7 CpG motif and methylation

Heredity is not maintained only by the genetic information encoded by DNA. Another component of it is epigenetic information. This kind of information is provided by the differential DNA methylation on cytosines of CpG, by the covalent modifications of DNA associated proteins and hence chromatin.

The link between cell status and genetic and epigenetic control is as it follows. DNA associated proteins that are mostly responsible for the epigenetic information transmission are histones and methylated CpG binding proteins. Histones, H2A, H2B, H3 and H4, form nucleosome core around which ~165 nt DNA wrap; and histones can be covalently modified. These covalent modifications include methylation, acetylation, phosphorylation and/or ubiquitination. The status of histones in terms of these modifications and methylation of DNA is linked to DNA accessibility together with the repertory and abundances of transcription factors and thus determines the transcription status of the genes. As a more clear statement, methylated DNA, methylated histones and compact chromatin structure define inaccessible DNA whereas acetylated histones and open chromatin structure means accessible DNA (Fitzpatrick and Wilson, 2003).

There are three DNA methyl transferases. The one called *Dnmt1* functions in the S phase of the cell cycle during which *Dnmt1* expression is upregulated. It is responsible for the methylation of CpG on newly synthesized DNA by taking the old strand as the reference. This implies that *Dnmt1* is required for methylation pattern maintenance (Jackson-Grusby et al., 2001; Knox et. al., 2000; Lee et. al., 2001; Li et. al., 1992; Robert et. al., 2003; Fitzpatrick and Wilson 2003).

Other DNA methyltransferases, *Dnmt3a* and *Dnmt3b*, are mainly expressed in embryonic development. They don't use hemi-methylated DNA as their substrates, yet they perform de-novo methylation. *Dnmt3L* is another member of

methylation repertory of mammals, which regulates the function of *Dnmt3a* and *Dnmt3b* by binding to them. (Bourc'his et. al., 2001; Chedin et. al., 2002; Hata et. al., 2002; Fitzpatrick and Wilson 2003).

These are the molecules that contribute to determine the fate of the gene expression. The mechanism can be in two different ways. First of all, methylation itself can block the binding of required transcription factors to the promoters. Secondly, methylated DNA binding proteins recruit histone deacetylases and other repressors to the site, making the locus inaccessible.

CHAPTER 2: AIMS

MicroRNAs are ~21-22 nt RNA sequences playing a switch role in the last transition part of central dogma, i.e., if they are in the scene, their existence ceases the protein formation. There are different kinds of bioinformatics studies that investigate expression and sequence information miRNAs hold; target prediction is one of the most studied. However, miRNA sequences have not been investigated in terms of their dinucleotide and trinucleotide frequencies.

Karlin and Gentles (2001) announced a fact boldly: dinucleotide frequencies differ in different genomes but this difference is significantly low between equally-sized pieces of the same genome. Accordingly, several questions emerge: (1) How are the dinucleotide frequency characteristics of human genome reflected on microRNA sequences? (2) Are there any over-represented and/or under-represented dinucleotide and/or trinucleotide motifs in both the mature and pre-mature microRNAs? In fact, the comparison of miRNA sequence features from the mature and pre-mature miRNAs could give interesting insights in terms of size-dependency of the microRNA specific dinucleotide and trinucleotide motifs.

Expression analysis through the microarray technology has been a great advancement in the recent years providing comprehensive information about the expression profiles of multiple genes simultaneously. Another aim of this study therefore was to test whether there is any correlation between the highly underrepresented CpG motif and the expression patterns of miRNAs in mammals, in particular the human and mouse models.

CHAPTER 3: METHODS

3.1 Extraction and reformatting of experimental and predicted pre-mature and mature miRNA sequences

First part of this study involved the nucleotide composition analysis of both the registered and predicted miRNA species in human. 222 pre-mature and 207 mature miRNA sequences have been fetched from Sanger miRNA registry (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl) and saved as text files Similarly, 976 newly predicted pre-mature microRNA sequences extracted from Berezikov et al. (2005) also were used in the analysis. All registered and predicted sequences were reformatted by replacing T with U.

3.2 Calculation of dinucleotide and trinucleotide frequencies

In order to calculate the dinucleotide and trinucleotide frequencies of the registered and predicted miRNAs, perl codes were used (*mir.pl* and *mirp.pl*, respectively; Appendix). The frequencies of dinucleotide and trinucleotide motifs were calculated for all possible frames (two for dinucleotide and three for trinucleotide); then each value was multiplied by 100 and logarithmically transformed (natural). Similarly, dinucleotide frequency values also were calculated without log transformation. The results for dinucleotide and trinucleotide frequencies were tabulated as matrices in which rows indicated miRNA sequences while columns specified the individual motifs. Matrices for each frame were written into separate text files; opened using Microsoft Excel for further analysis; and an average value for each motif (column) was calculated.

3.3 Estimating the expected frequencies using a randomization procedure

The miRNA sequences were shuffled individually using a perl code (*shuffle.pl*; Appendix). The average dinucleotide and trinucleotide motif frequencies of the resulting sequences were calculated based on methods described Section 2.2; and the resulting randomized average nucleotide frequencies were saved as text files. Sequentially, the sequences were randomized; individual dinucleotide frequencies averaged; and this procedure is repeated multiple times (N = 1000, *newshuffle.pl*; Appendix). The resulting output included the randomized dinucleotide means as rows, which then averaged using Microsoft© Excel. Next, the '*pvalue.pl*' (Appendix) was used to calculate the rank of the observed dinucleotide frequency of a particular motif within the set of expected values generated, which provided the probability of having a significantly deviating nucleotide frequency.

3.4 Number of miRNAs without a certain dinucleotide or trinucleotide motif

A perl code (*mcount.pl*; Appendix) was written to detect the presence or absence of each of the 16 dinucleotide motifs in the miRNA sequences and to generate a binary text file as the output. The frequency distribution of miRNA sequences with or without a particular motif was determined by counting the number of absences.

3.5 Estimation of the asymmetry in the nucleotide occurence at the 5' and 3' of miRNAs

The first two bases of each end of a mature[3] microRNA were determined and tabulated (*first.pl* and *last.pl*; Appendix). These programs was used to create 16 headers, one for each dinucleotide motifs; and to record and count the microRNAs that have the particular motif indicated by the header, at the first or last position.

[3] The mature sequences of predicted microRNAs are not reported.

To further investigate the polarity issue between 5' and 3' ends of registered microRNAs, a perl code '*vectors.pl*', which generated a matrix (207 x 32) to represent the existence of any of the four bases at the 3' and 5' ends also was written (Appendix). Furthermore, all pairwise jaccard similarity indices betweens the vectors were calculated to estimate the similarity between the base content of the 3' and 5' ends of the mature miRNA sequences (*jaccard.m*; Appendix).

Jaccard coefficient calculates how two equally-sized column vectors of zeros and ones close to each other. The formula is given below:

$$J=C/(C+N1+N2)$$

where *J* is the jaccard coefficient, *C* is the number of common ones, *N1* is the number of differences in vector one and *N2* is the number of differences in vector two.

3.6 Extraction and formatting of the mature human miRNA sequences from the registry

The mature sequences of the microRNAs were downloaded from Sanger miRNA registry (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl) and saved as text files. The purpose was to use the most updated list of microRNAs in expression analysis. This text file was modified via Microsoft© Excel 2003 such that the RNA letters were replaced with capital DNA letters.

3.7 The expression data

Baskerville et al., (2005) reported the expression profiles of 175 human microRNAs. The expression data were downloaded from the following site:

http://web.wi.mit.edu/bartel/pub/bartel_publications.html.

The downloaded supplementary information reported the observed the expression levels of 175 human along with those of other organisms in 25 different tissues. Expression data from the *H. sapiens* was extracted and saved as a text file.

An independent dataset focusing on the mouse embryonic and adult miRNA expression pattern (Thomson et al., 2004) also was used in this study. The expression profiles, RNA sequences and list of the 124 investigated murine microRNAs were downloaded from the following supplementary web address:

http://www.nature.com/nmeth/journal/v1/n1/suppinfo/nmeth704_S1.html

The embryonic and tissue-specific expression data provided by Thomson et al. (2004) were separately recorded; the microRNA names together with their sequences were saved in FASTA format.

3.8 Extraction of microRNAs common in registry and human microarray data

A perl code (*cmirs.pl*; Appendix) was written to extract and save the names of microRNAs common in the human expression data and registry record (Appendix). A second perl code (*cmirs2.pl*; Appendix) accomplished a similar task to produce a text file that contained expression values together with the corresponding miRNA names. Briefly, microRNA names common to both datasets were compared to those in the expression data one by one; and when a match occurred, the data were recorded.

3.9 Calculation of mean expression levels for miRNAs with or without a CpG motif

After the determination of common microRNAs, miRNA species were separated based on the presence or absence of a CpG motif in them and associated with the corresponding expression data using *cpgn.pl* and *cpg.pl* (Appendix) for both the human and mouse data.

Another perl code (*cpgmean.pl*) was implemented for the calculation of tissue-specific mean expression values. A mean expression value was obtained for each of the tissues. Any miRNA, with a missing expression data (indicated by an NaN), was not included in the calculation of the mean expression value.

The perl generated results were compared with the results derived from the Microsoft©Excel to confirm the accuracy of the calculations. All the procedures described in this section were applied to data derived from the study done by Thomson et al., (2004) in a tissue source discriminating manner.

3.10 Randomization

The mean values of expression levels of randomly selected non-CpG microRNAs were calculated using *randomly.pl* (Appendix). The number of selection was fixed and equaled to the number of CpG containing microRNAs. The mean values were calculated for 10000, 20000 and 50000 randomization steps for raw, log normalized, and log normalized median subtracted datasets, separately. Randomization was accomplished using the *shuffle*[4] function provided by the Active State Perl Version 5.8+, which produced a different permutation of a list or of an array, each time. An array that equaled the length of the nonCpG microRNA was produced. Elements of this array reflected the indices of the array. The mean expression value of the nonCpG microRNAs called by each randomized index was calculated and saved as a text file.

The standard deviations on the expression values published by Baskerville et al. (2005) were relatively large. Therefore, this existing variation was normalized in two different ways for a more robust analysis. As a simple approach, the logarithms of expression values were taken. As a second normalization method, the expression values were divided by a column-specific median for each tissue and then logarithms of theses values were calculated. These calculations were made via implementation of changes in the '*randomly.pl*'. Data from the mouse study (Thomson et al., 2004) also was analyzed using the same procedure while embryonic and adult expression data were treated, separately.

[4] For a detailed discussion of the issue: http://perl.active-venture.com/pod/perlfaq4-dataarrays.html

3.11 Determination of the Significance of Expression Differences

After the production of a set of mean values obtained from randomly selected non-CpG miRNAs, the number of random mean values maller than those of the CpG containing microRNAs was calculated for each tissue separately; and these values were divided by the number of randomization steps. This task was accomplished by the perl implementation '*probable.pl*' (Appendix).

CHAPTER 4: RESULTS

Dinucleotide frequencies for the premature miRNA sequences, calculated using either frame, indicated that CpG dinucleotide frequency was relatively under-represented when compared with other dinucleotide motifs both for the registered and predicted premature miRNAs (Figure 4.1, Figure 4.2).

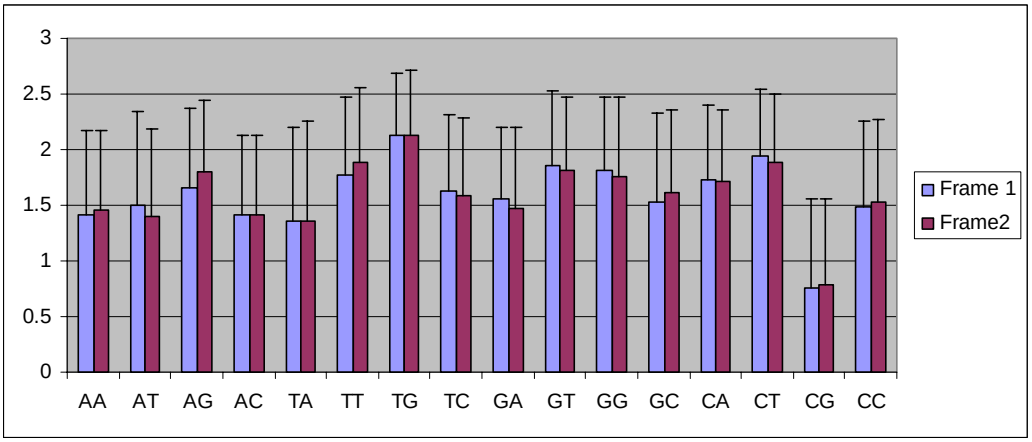

Figure 4.1 Average dinucleotide frequences in the pre-mature micro RNA sequences.

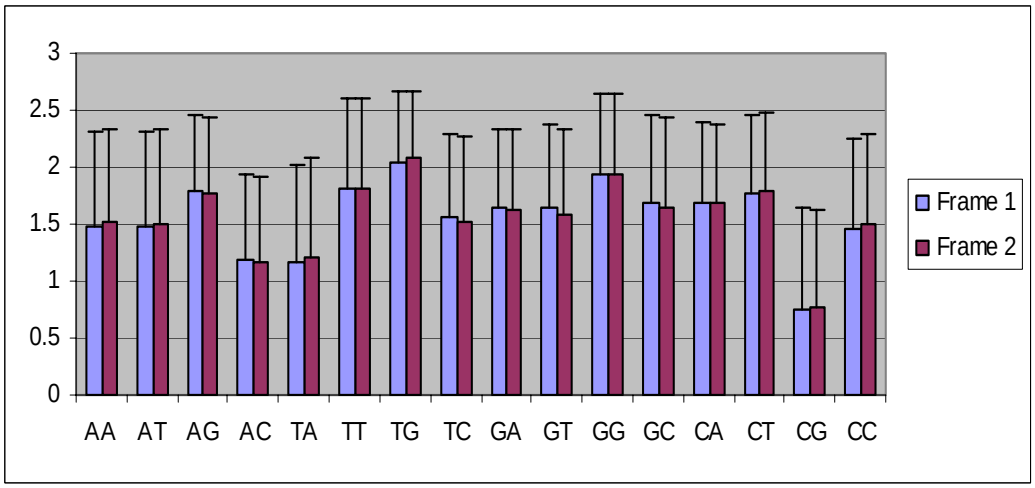

Figure 4.2 Average dinucleotide frequencies of the pre-mature miRNA sequences predicted by Berezikov et al (2005).

Randomized miRNA sequences, on the other hand, produced average dinucleotide frequencies, which were more uniform in distribution (Figure 4.3 and Figure 4.4, for registered and predicted sequences, respectively).

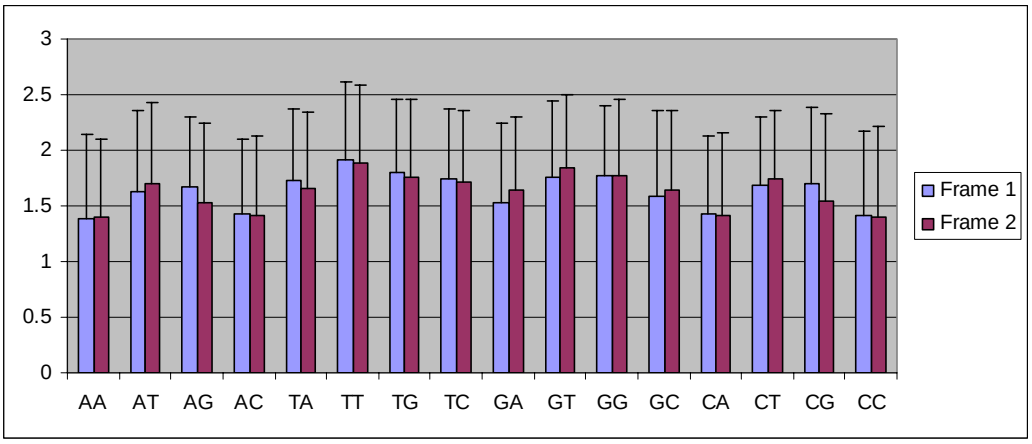

Figure 4.3 Average dinucleotide frequencies of the shuffled pre-mature miRNA sequences obtained from Sanger registry.

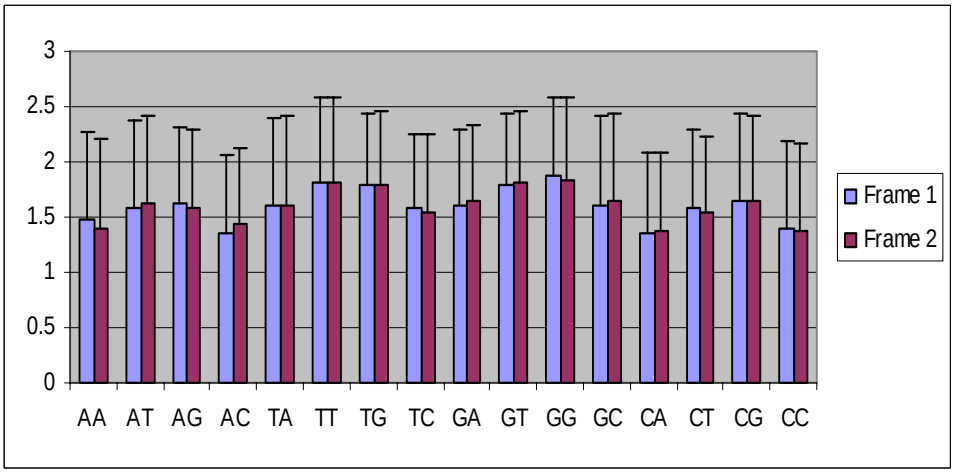

Figure 4.4 Average dinucleotide frequencies of the shuffled miRNA sequences predicted by Berezikov et al (2005).

A similar schema also emerged for the mature sequences (Figure 4.5) such that CpG dinucleotide was the least frequent motif whereas over representation of TpG dinucleotide appeared to be distinctive.

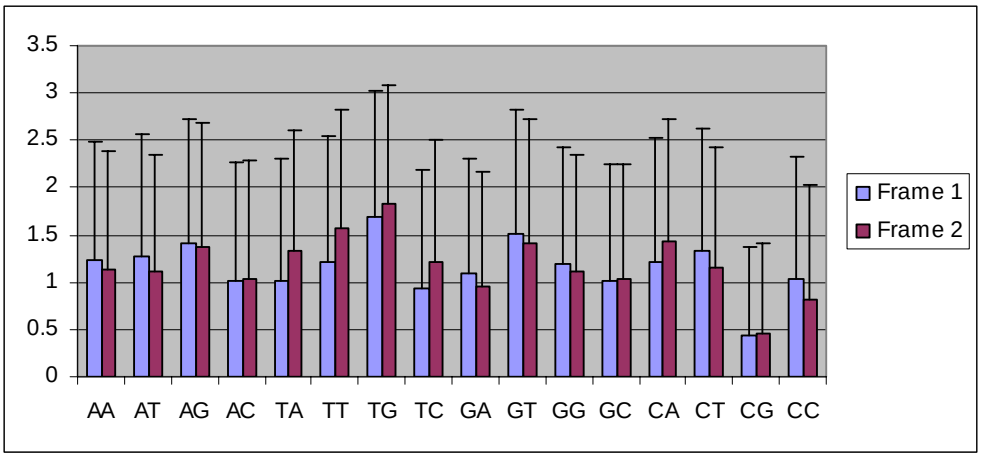

Figure 4.5 Average dinucleotide frequencies of the mature miRNA sequences registered in Sanger database.

The number of registered microRNAs that did not contain a particular dicnuleotide motif also was investigated (Figure 4.6). This graph demonstrated that the number of mature miRNAs lacking CpG and CpC motifs were strikingly high whereas TpG dinucleotide was frequently present in many mature miRNAs such that only 20 out of 207 mature sequences did not contain TpG (Figure 4.6). Although we observed similar results in the candidate miRNA sequences predicted by Berezikov et al, (2005), these results can be biased due to the large amount of similar microRNAs that belonged to the same mir families.

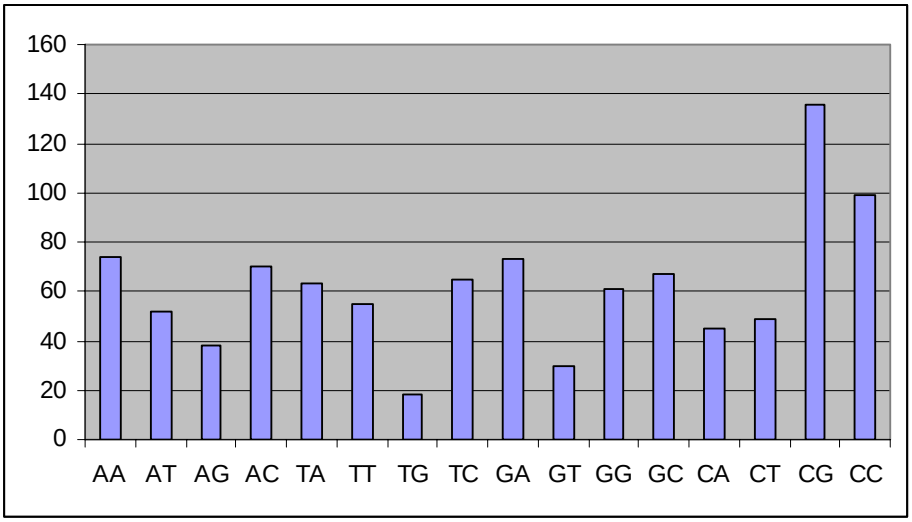

Figure 4.6 The number of microRNAs that don't contain a certain dinucleotide motif.

Randomizations were performed to estimate the statistically significant over- and under-represented dinucleotide motifs for the collection of registered mature miRNAs (log transformed; Figure 4.7 and Table 4.1). The results indicated that AG, TG, CA, and CG motifs were highly significantly different between the observed and randomized average dinucleotide frequency values ($p < 0.001$ based on 1000 randomizations). The untransformed data resulted in comparable p-values for these four motifs although the overall extent of the significance was reduced (data not shown).

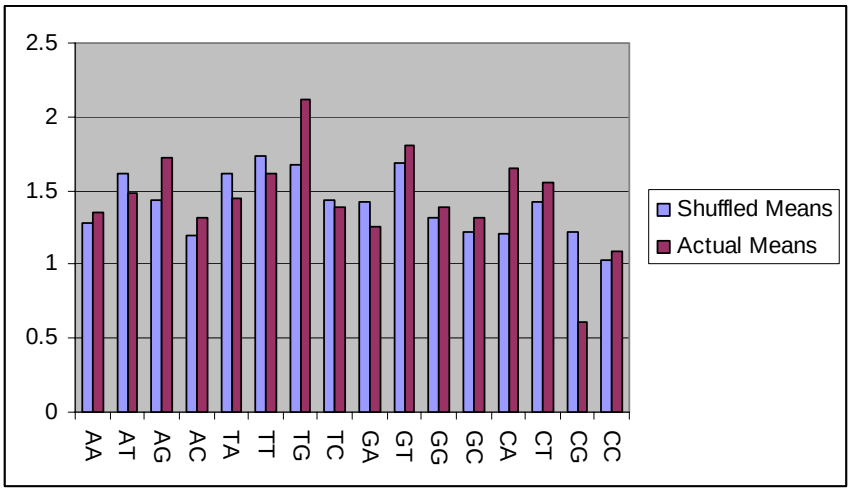

Figure 4.7 The comparison of actıal average dinucleotide frequency with thosefrom the randomized set of mature sequences.

Table 4.1 The probability of the observed dinucleotide frequency not being different from those obtained for the randomized miRNA sequences (N = 1000).

Dinucleotide	Probability
AA	0.075
AT	0.014
AG	0
AC	0.025
TA	0.001
TT	0.006
TG	0
TC	0.22
GA	0.005
GT	0.023
GG	0.11
GC	0.06
CA	0
CT	0.012
CG	0
CC	0.155

In a similar manner as performed for dinucleotide motif content analysis of all microRNA sequences, the trinucleotide motifs also were analyzed.

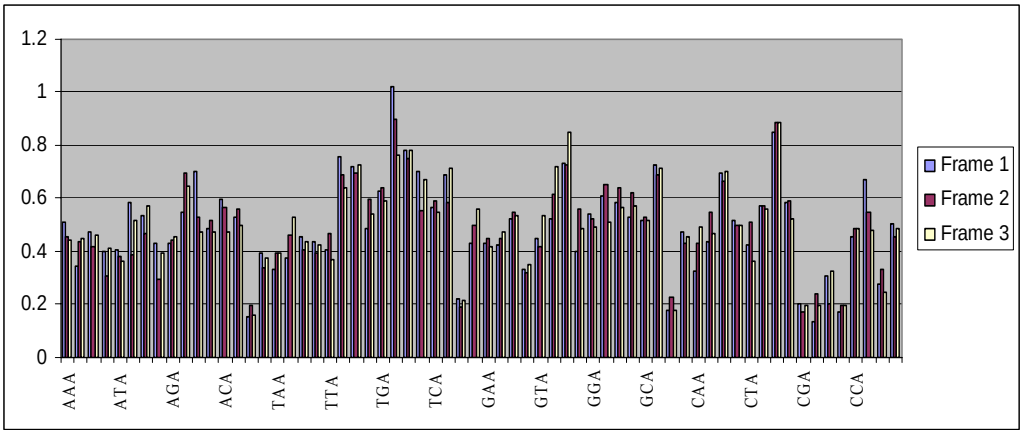

Figure 4.8 Average trinucleotide content of the registered pre-mature miRNAs.

As seen from the Figures 4.8 and 4.9 the under-representation of CpG containing motifs and over-representation of TpG containing trinucleotide motifs were predominant features of the premature and mature miRNA sequences.

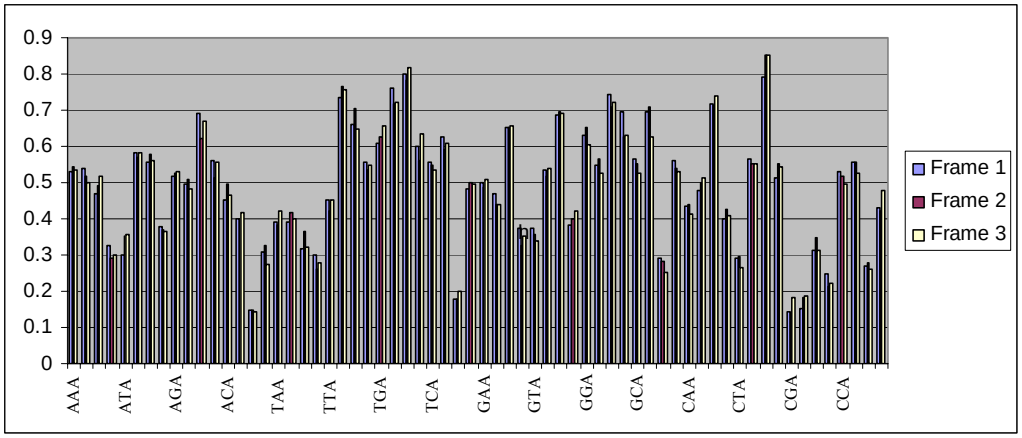

Figure 4.9 Average trinucleotide content of the predicted mature miRNAs.

Another hypothesis to be tested was whether there was any polarity between the 5' and 3' ends of the microRNA sequences in terms of the choice of base composition.

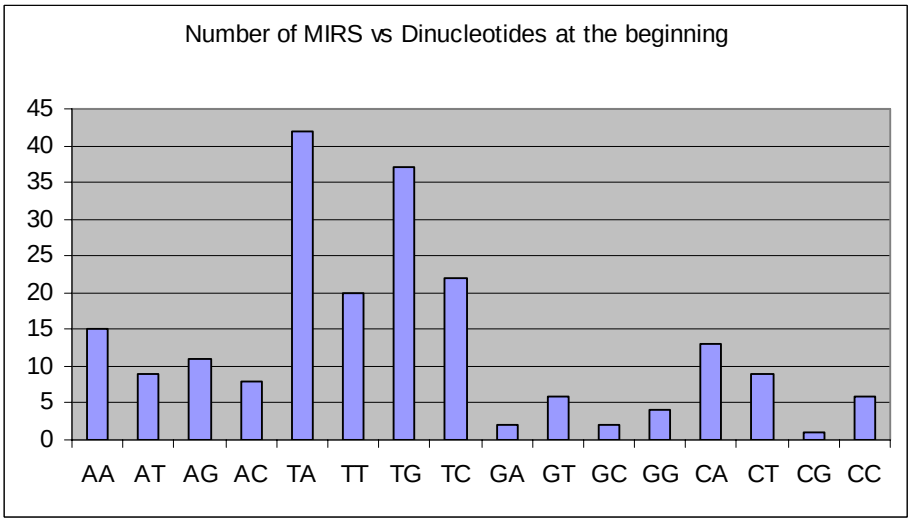

Figure 4.10 Number of miRNAs with a particular dinucleotide motif at the 5' end.

Mature microRNAs were found to have a T nucleotide in their first 5' position frequently while G nucleotide was low in frequency at the same position (Figure 4.10). Additionally there was only one microRNA, *hsa-miR-324-5p*, that started with CpG dinucleotide motif.

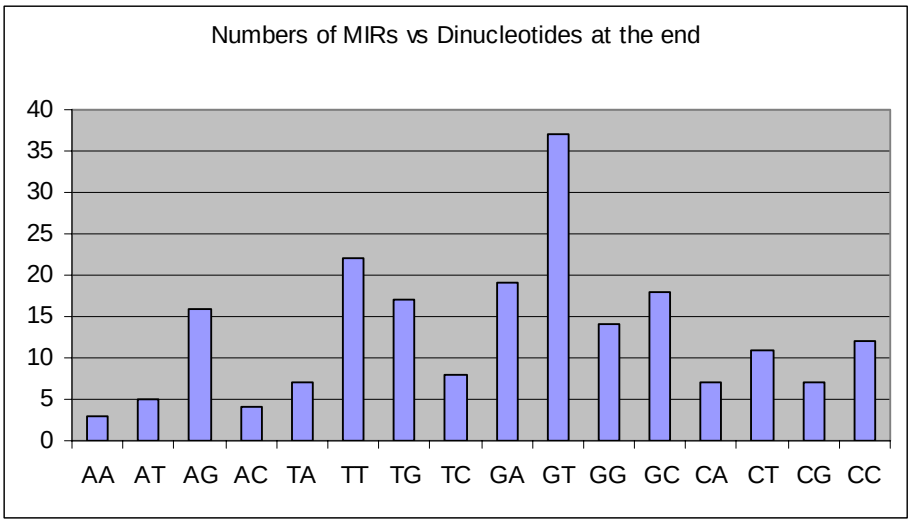

Figure 4.11 Number of miRNAs with a particular dinucleotide motif at the 3' end.

However, there seemed to be less preference for a specific base at the 3' end when compared with what was observed at the 5'end (Figure 4.11). However, the G nucleotide, which was not frequently seen at the 5' end, was rather over-represented at the 3' end.

Table 4.2 Pairwise Jaccard coefficients between base vectors (the first position from either the 5' or the 3') obtained for the set of registered mature miRNA sequences.

	5'T	**G**	**C**	**A**	**3'T**	**G**	**C**	**A**
5'T	1	0	0	0	0.2763	0.2429	0.1408	0.1818
G	0	1	0	0	0.0602	0.0968	0.0182	0.0417
C	0	0	1	0	0.1196	0.0506	0.1639	0.0656
A	0	0	0	1	0.1584	0.1149	0.1486	0.0822
3'T	0.2763	0.0602	0.1196	0.1584	1	0	0	0
G	0.2429	0.0968	0.0506	0.1149	0	1	0	0
C	0.1408	0.0182	0.1639	0.1486	0	0	1	0
A	0.1818	0.0417	0.0656	0.0822	0	0	0	1

According to the Jaccard's coefficient matrix, if the nucleotide in the 5' end was a T, the nucleotide in the 3' end was equally likely to be a T or a G (Table 4.2). On the other hand, if the nucleotide in the 3' end was a G, the likelihood of having a T was greater than any other. Nevertheless, a similar bias was not apparent with respect to A or C nucleotides. Information obtained from the Jaccard coefficients, although provided potentially useful information about the potential assymetry on the

selection of A-T or G-C base pairing at the 5' and 3' ends respectively; is likely to be biased because the jaccard coefficient is based on the number of co-occurences, and thus would be highly affected by the nature of presence/absence data. Since certain nucleotides were under-represented in the collection of the miRNAs, this might have affected the value of the jaccard's index.

The correlation between the under-representation of the CpG motif and the expression value of the tissue- and stage-specific miRNAs also was investigated. The comparison of 222 updated registry mature microRNAs to the microRNAs that were investigated in the expression study by Baskerville and Bartel (2005) resulted in a list of 148 microRNAs that were common to both studies, and 50 of which contained at least one CpG dinucleotide motif.

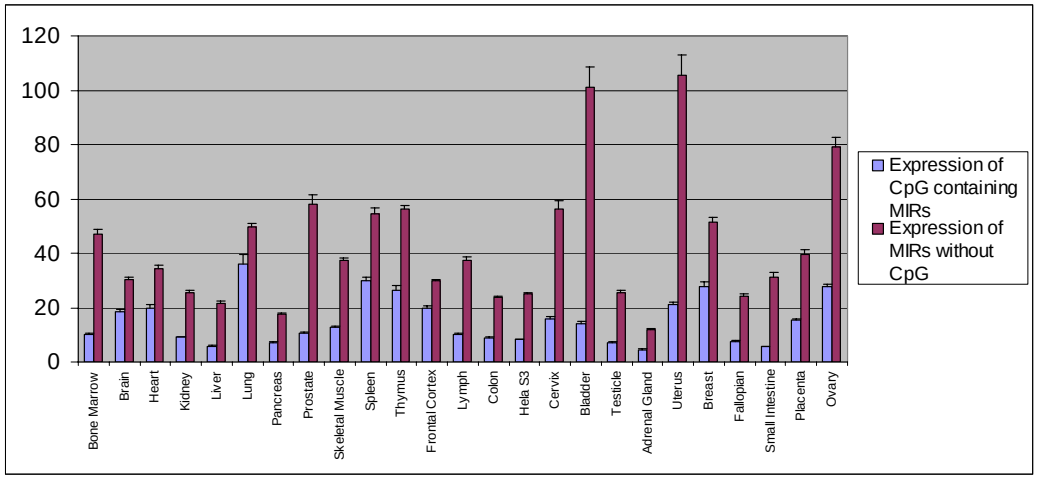

Figure 4.12 Bar graphs of mean expression values of miRNAs with or without a CpG motif in human, based on the raw expression data provided by Baskerville and Bartel (2005).

Mean differences between expression signals derived from the two groups of miRNAs (CpG and no-CpG) were found to be striking. Expression signals from the microRNAs bearing CpG dinucleotide motif was lower than the ones obtained from the no-CpG group (Figure 4.12).

Randomizations were performed to test whether these differences were significant (Table 4.3). Three different randomization steps were used to estimate the optimum number of randomization required (i.e., 10000, 20000 and 50000, respectively). It was noticable that the probability values did not change drastically

with the increased number of iterations, indicating that even the 10000 iterations was adequate to explain the variability in the expression dataset (Table 4.3).

Table 4.3 The mean expression values for the miRNA with or without CpG motif for raw data and the significance of the difference based on randomizations, (i.e., 10000, 20000, 50000).

	EXPRESSION LEVEL				
TISSUES	**No CpG**	**CpG**	**P(10000 iter)**	**P(20000 iter)**	**P(50000 iter)**
Bone Marrow	47.23	10.32	**0**	**0.00012**	**0.00004**
Brain	30.36	18.61	0.0248	0.02518	0.02658
Heart	34.25	19.66	0.0831	0.08778	0.0865
Kidney	25.57	9.14	**0**	**0**	**0.00002**
Liver	21.63	5.9	0	0	0
Lung	49.48	35.99	0.1828	0.1903	0.1842
Pancreas	17.78	7.22	**0.0001**	**0**	**0**
Prostate	58.13	10.65	0.0004	0.00016	0.00006
Skeletal Muscle	37.46	12.82	0	0.00002	0.00002
Spleen	54.7	29.68	0.1146	0.1209	0.11908
Thymus	56.07	26.33	0.0056	0.00726	0.0077
Frontal Cortex	29.78	19.82	0.043	0.04714	0.04586
Lymph	37.58	10.17	**0.0005**	**0.00032**	**0.0002**
Colon	23.58	8.87	**0**	**0**	**0**
Hela S3	24.84	8.15	**0**	**0**	**0.00002**
Cervix	56.48	15.82	0.0028	0.00208	0.00224
Bladder	101.21	14.14	0.0022	0.00136	0.00148
Testicle	25.47	7.13	**0.0001**	**0**	**0.00002**
Adrenal Gland	11.8	4.61	0.0009	0.00056	0.0007
Uterus	105.67	21.08	0.01	0.0127	0.01208
Breast	51.39	27.91	0.036	0.03926	0.0393
Fallopian	24.36	7.56	0	0.00002	0.00002
Small Intestine	31.28	5.67	**0**	**0.00002**	**0**
Placenta	39.62	15.46	0.0194	0.01914	0.01978
Ovary	79.19	27.76	0.0014	0.00158	0.00148

When the data were log-normalized, the probability of mean expression difference between CpG or noCpG miRNAs (Table 4.4) exhibited differences from results shown in Table 4.3.

Table 4.4 The log normalized mean expression values for the miRNA with or without CpG motif and the significance of the difference based on randomizations (N = 10000).

	EXPRESSION LEVEL		
TISSUES	**No CpG**	**CpG**	**Probability**
Bone Marrow	1.835641	1.09907	0.0001
Brain	2.272933	2.04823	0.0397
Heart	2.244487	1.81768	0.0008
Kidney	2.048454	1.37289	0
Liver	1.973315	1.37383	0
Lung	2.235088	1.45285	0.0003
Pancreas	1.852805	1.04422	0
Prostate	1.796364	1.1949	0
Skeletal Muscle	2.203071	1.68997	0.0007
Spleen	2.293949	1.6963	0.0006
Thymus	2.406602	1.81433	0.0003
Frontal Cortex	1.971126	1.86395	0.2117
Lymph	1.933343	1.32089	0
Colon	2.176913	1.46098	0
Hela S3	1.840348	1.21463	0.0001
Cervix	1.978211	1.37922	0.0006
Bladder	1.876765	1.23022	0.0004
Testicle	1.682284	1.00768	0
Adrenal Gland	1.133019	0.667	0.0002
Uterus	2.016614	1.45267	0.0017
Breast	2.243979	1.79119	0.0075
Fallopian	1.767602	0.94277	0
Small Intestine	1.572954	0.91893	0
Placenta	2.052875	1.60639	0.0018
Ovary	2.633092	2.04853	0.0006

.

Earlier, it has been demonstrated that the diferences between probabilities after 10000 and 50000 steps of randomizations the probabilities were comparable. Accordingly, the number of the randomization step for the log normalized data sets was choosen as 10000. Upon logarithmic transformation of the human expression data, only the Brain and Frontal cortex regions preserved their insignificant status. Expectedly, the results of log-normalized and log-normalized and median subtracted

data were highly similar (Table 4.5). The persisting relatively high probabilities from the brain and frontal cortex were again striking.

Table 4.5 The median subtracted log normalized mean expression values for the miRNA with or without CpG motif and the significance of the difference based on randomizations (N = 10000).

	EXPRESSION LEVEL		
TISSUES	**No CpG**	**CpG**	**Probability**
Bone Marrow	0.83	0.09	0
Brain	0.71	0.47	0.0386
Heart	1.08	0.65	0.0005
Kidney	0.66	0.04	0
Liver	0.98	0.4	0
Lung	0.52	-0.26	0
Pancreas	0.65	-0.16	0
Prostate	0.65	0.06	0.0002
Skeletal Muscle	1.24	0.74	0.0003
Spleen	0.54	-0.06	0.0001
Thymus	0.75	0.19	0
Frontal Cortex	0.63	0.53	0.2654
Lymph	0.95	0.32	0
Colon	0.45	-0.26	0
Hela S3	0.63	0.04	0.0003
Cervix	0.63	0.03	0.0004
Bladder	0.53	-0.11	0.0004
Testicle	0.47	-0.21	0
Adrenal Gland	0.75	0.29	0.0004
Uterus	0.75	0.17	0.001
Breast	0.54	0.09	0.0081
Fallopian	0.55	-0.27	0
Small Intestine	0.71	0.06	0
Placenta	0.66	0.21	0.0017
Ovary	0.6	0.01	0.0005

The mean values for the median normalized expression levels of CpG containing microRNAs and non-CpG microRNAs also were graphically illustrated in Figure 4.13. In all of the tissues but brain and frontal cortex, the CpG containing microRNAs were expressed in higher amounts than the ones that do not contain any CpG.

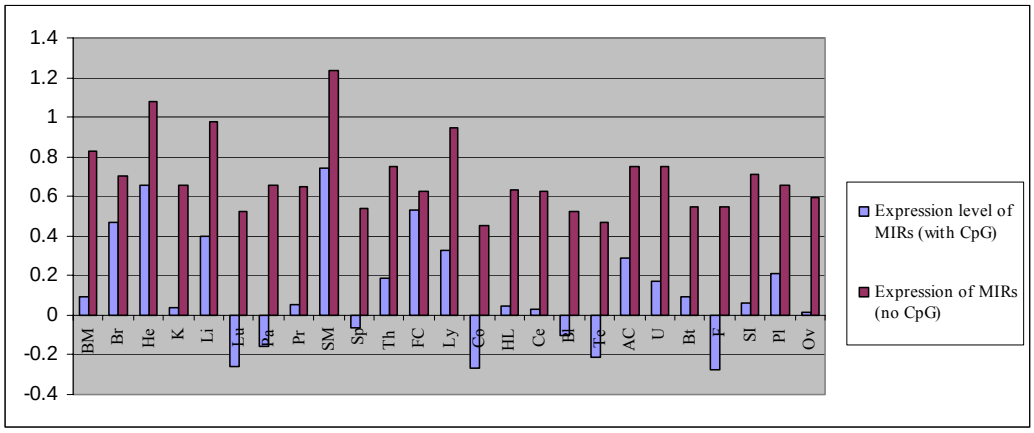

Figure 4.13 Bar graphs of median subtracted log normalized expression values of miRNAs with or without a CpG motif in human.

In order to obtain further support for the findings shown in Tables 4.3-4.5, another microarray dataset, in this case from the mouse, also was investigated as stated in Chapter 2 (section 2.8). The expression of common mammalian microRNAs had been reported from seven different mouse tissues and cells from seven different embryonic stages from mouse embryo. This dataset was already log-normalized therefore was used as is. The mean differences between the CpG and nonCpG miRNAs were strikingly different except for the brain and to a lesser degree in thymus (Table 4.6, Figure 4.14).

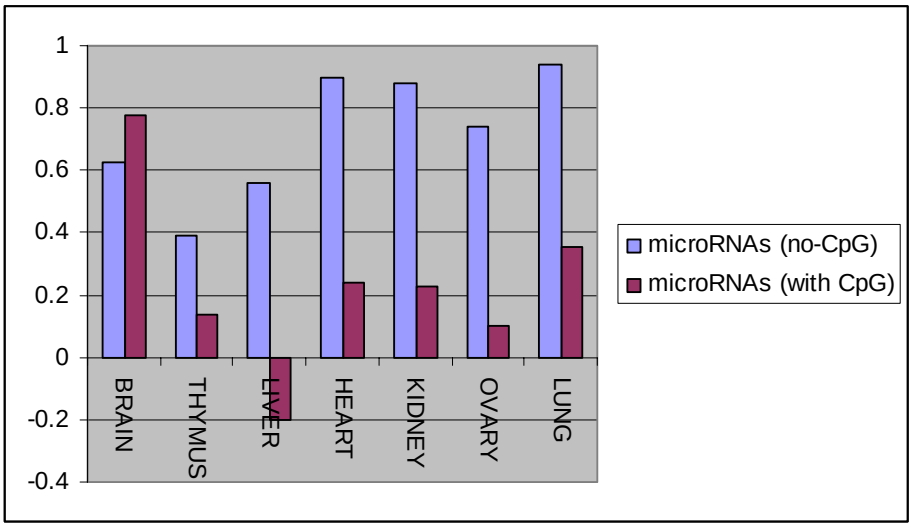

Figure 4.14 Tissue-specific differences in terms of the mean normalized expression level for CpG and nonCpG miRNAs among mouse tissues.

Table 4.6 The probability of having a higher expression signal for the CpG containing microRNAs in mouse tissues.

EXPRESSION LEVEL	
TISSUES	**Probability**
BRAIN	0.7225
THYMUS	0.052
LIVER	0
HEART	0.001
KIDNEY	0.0002
OVARY	0.0002
LUNG	0.0017

Interestingly, the results from mouse tissues were comparable with those from human counterparts (Table 4.7). In both species, the brain was the region where no expression difference between CpG and non-CpG miRNAs. Low probabilities calculated from liver and kidney were very significant in both organisms. All other tissues also were significantly different in terms of average expression between the miRNAs with and without CpG (Table 4.7).

Table 4.7 Comparison[5] of probabilities derived from human data (Table 4.4) and mouse data (Table 4.6) at 10000 randomization level.

Human				Mouse	
TISSUES	No CpG	CpG	Probability	TISSUES	Probability
Brain	2.27293	2.04823	0.0397	**Brain**	0.7225
Thymus	2.4066	1.81433	0.0003	**Thymus**	0.052
Liver	1.97332	1.37383	0	**Liver**	0
Heart	2.24449	1.81768	0.0008	**Heart**	0.001
Kidney	2.04845	1.37289	0	**Kidney**	0.0002
Ovary	2.63309	2.04853	0.0006	**Ovary**	0.0002
Lung	2.23509	1.45285	0.0003	**Lung**	0.0017

The expression pattern for the two types of microRNAs in mouse embryonic tissues gave a relatively different profile than what was observed in the adult tissues (Figure 4.15)

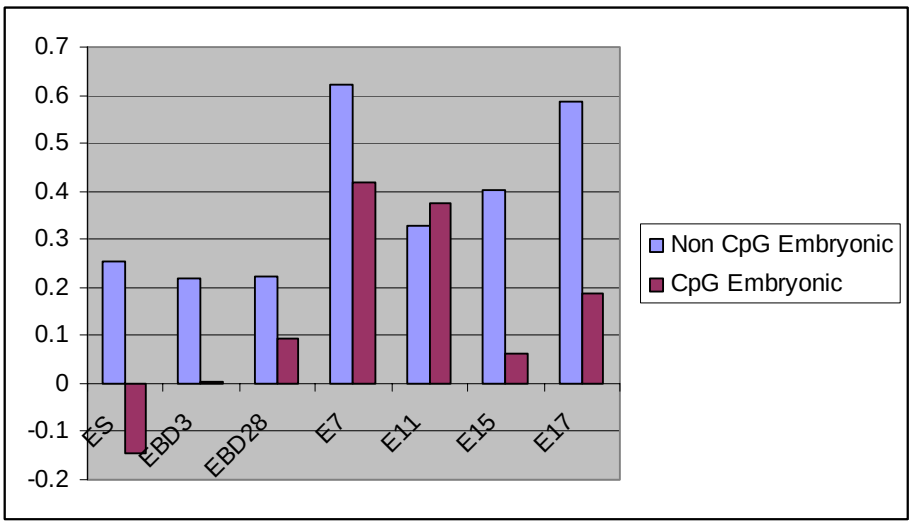

Figure 4.15 CpG containing microRNAs are compared with the CpG free ones in cells of the mouse embryo of different stages in terms of expression level.

The embryonic samples in this dataset could be separated into two groups, one of which was the embryonic stems cells and embryoid bodies while the other group of embryos represented the different stages of embryonic development from E7 to E17. The mean expression level of both the CpG and nonCpG miRNAs

[5] Human data as reported in original paper

increased as the embryo aged while the expression of CpG containing miRNAs were significantly underrepresented in the stem cells and embryonic bodies as well as the later stages of the embryonic development, i.e., E15 and E17 (Table 4.8).

Table 4.8 The chance probability of having a higher expression signal for the CpG containing microRNAs in embryonic mouse cells (10000 randomizations).

EXPRESSION LEVEL	
TISSUES	**Probability**
ES	0
EBD3	0.008
EBD2	0.081
E7	0.107
E11	0.611
E15	0.005
E17	0.005

CHAPTER 5: DISCUSSION

Several intriguing patterns have become apparent upon the analysis of dinucleotide frequencies of the registered and predicted miRNAs. Particularly interesting was that the registered microRNA sequences were scarce in terms of the CpG motif while they contained the TpA as the second most frequent dinucleotide motif. Predicted microRNA sequences had similar dinucleotide contents to the registered ones. While there was no specific polarity between the ends of mature miRNA squences, it could be said that the 5' end of the mature microRNA sequences started with T preferentially. Interestingly, the G nucleotide has been observed so rare at the 5' end but the reverse was correct for the 3' end. Correlation of the expression levels of mature microRNAs with the presence/absence of the CpG motif revealed that mature microRNAs with CpG motif were expressed in significantly lower amounts when compared to mature microRNAs without the CpG motif, in mouse and human.

After the determination of full genome sequences for various species, the wonder of knowing compositional differences between them also have lead to complementary researches. One of the most interesting studies was performed by Gentles and Karlin (2001): They have investigated the dinucleotide frequency differences across the species. Relative dinucleotide frequency abundances for each genome generally are unique while the variability within the genome is low. According to Gentles and Karlin (2001), the main characteristic of human genome in terms of dinucleotide frequency abundances was found to be the relatively high frequency of CC/GG, AG/CT and CA/TG dinucleotide motifs. The other hallmarks were the under-representation of the CpG motif, a shared property with *Mus musculus* genome, and the relatively low abundance of TA dinucleotide motif, common in all organisms tested[6].

In the present thesis, I have focused on answering a scientific curiosity involving the compositional pattern of the dinucleotide frequencies for human

---

;[6] *Human, M. musculus, C. elegans, D. melanogaster, S. cereviisae, A. thaliana, Plasmodium falciparum and Leishmania major*.

microRNA sequences as it relates to the whole genomic signature. Accordingly, my question was how Gentles and Karlin's report fits to these interesting small non-coding sequences?

At first glance, CpG dinucleotide under-representation observed in the whole human genome could be extended to the miRNA sequences studied herein. The significance of this finding was clear after comparing the actual miRNA sequence composition with their shuffled counterparts without changing their base contents. Shuffling has removed this under representation making it obvious that CpG was very highly significantly underrepresented in the registered and predicted human miRNAs.

The number of mature microRNAs that begin and end with a certain dinucleotide has been also studied in this study. It appeared that, neither at the beginning nor at the end of a sequence was there a single predominating dinucleotide motif. But it was striking to find that although the frequency of TA dinucleotide was not the highest one, relatively speaking; the number of microRNAs that begin with TA appeared to be the highest among all the other fifteen dinucleotide motifs (Figure 4.10). Even any preference can not be claimed for dinucleotides at the 5' end of the mature sequences, at the single nucleotide level, a claim of overrepresentation of T at the 5' was considerably likely. Seemingly, it can be said that there was a G preference at the 3' end (Figure 4.11). In order to be functional, the double stranded form of microRNAs must be separated. Hence when RISC complex is associated with the double stranded species, most of the times the T-A bonds are broken at 5' end. It may be energetically preferential to possess a T for the cell at the 5' end of its mature microRNAs.

When the comparison issue is considered for mature sequences, it cannot be said that TA is the second-most underrepresented motif. Although it may not be favored in the messenger RNAs because of the vulnerability of UA to ribonucleases (Beutler et al., 1989), in mature microRNA sequences, TA did not seem to be underrepresented. This may be perceived as that the vulnerability of the mature sequences to ribonucleases may not be a problem for a cell. In addition, mammalian microRNAs play their regulatory roles by translation block, frequently without

degrading the messenger RNA[7]. Hence, mRNAs blocked by microRNAs probably wait to be translated on time. At this point the speculated vulnerability may serve for re-activating the translation of messenger RNAs by consuming free microRNAs in the cytoplasm, if we consider that transcription of unnecessary microRNAs mustn't occur. This may help explain how temporal regulation of developmental events occurs in the cell by microRNAs.

Methylation of the cytosine in CpG dinucleotide occurs in a wide variety of species. Methylation has been suggested to be the reason for under representation of CpG dinucleotides in genomes that bear methyltransferases, because deamination of methyl transferred cytosines to thymine nucleotide is a spontaneous event. As a result, there is a high likelihood for TpG in CpG deficient genomes. Figure 4.1 and Figure 4.5 seem to tell a similar scenario. In both mature and pre-mature sequences, TpG is the most frequent dinucleotide. Although they are relatively short sequences and CpG is under-represented throughout the genome. Figures 4.5-4.7 also showed that relatively less number of miRNAs contained a CpG motif; only about 50 mature microRNA sequences bear this motif whereas only 19 out of 207 mature sequences don't contain any TpG.

If methyl transferases still exist and are functionally conserved during evolution, the reason may be that they must be essential for organisms that bear it. Most likely, they allow the dilution of the CpG dinucleotides in the genomes. In molecular mechanisms that constitute life, the best candidate elements for rare and important regulatory events are the ones that are scarce in the media (i.e., Ca++). Hence, DNA methyl transferases may create a situation that can be used for regulating some events important via regulation of active CpGs content thus its accessibility.

For example, methylation of C residue of CpG dinucleotide at regulatory regions results in transcriptional repression. As another example, via methylated C binding proteins chromatin modeling complexes are recruited to the binding region making that region inaccessible and tightly packaged. Packaging of the DNA via

[7] microRNAs controlling the HOX gene clusters point the messenger RNAs to degradation.

methylated C signals largely determine the patterns for each cell type; and each pattern of this kind is heritable to progeny cells by the capability of DNA methyl transferase 1 (*Dnmt1*) methylating the sites on newly synthesized DNA exactly at the same positions. Hence methylation serves for keeping the identities of progeny cells same with their ancestors.

Apart from methylations, oligo DNAs and RNAs that contain CpG motif are perceived as foreign nucleic acids and innate immune response might be activated against them, helping protect from invasions by harmful microorganisms (Section 1.6).

In the light of implications in regards to the scarce CpG in human genome, what can be said about CpG dinucleotides in mature microRNA sequences? Although there are a significant number of microRNAs that bear CpG dinucleotide, the frequency is low. Several intriguing questions arise: Would there be any selective pressure on CpG containing miRNAs? Would this scarce CpG usage be as beneficial as they seem to be when the whole genome considered?

Without a wet-lab work it is hard to answer these questions in high certainty but a bioinformatics study can provide clues about the founding correlates. In order to get an insight into the underlying mechanism, I have compared the expression levels of CpG containing and CpG-free microRNAs.

The accuracy and precision of the quantification of expressed microRNAs is still debated, however, current methods still prove to be invaluable as evaluation tools. The mean expression levels calculated based on the Baskerville et al. (2005) data have shown that in human tissues, the expression of CpG containing microRNAs was lower than that of CpG-free micro RNAs, on average (Figures 4.12 and 4.13). The chance probability of the mean expression of CpG containing microRNAs being higher than the mean of a randomly selected pool of CpG free microRNAs was calculated for each tissue (minimum of 10000 iterations). These expected probabilities were very low for most of the tissues indicating significance (Tables 4.3-4.5). The tissues of exception with insignificant probabilities were the brain, heart, lung, spleen, frontal cortex, breast and placenta when raw (non-log

transformed) data were used. This shows a clear sign of a preference towards the usage of CpG-free microRNAs in the cells of different human tissues. A speculation trying to explain the differences in the probabilities between tissues can not be made easily, however, the relatively large probabilities suggesting insignificance of CpG motif as a correlate of expression in placenta, brain and frontal cortex tissues may indicate an evolutionary mechanism for escaping from self-immune attacks. Evidence that make this speculation supportable are related to some properties of immune system. First of all, it has been shown that CpG containing oligo nucleotides as mentioned in s*ection 1.6* can act as immune-stimulatory agents**.** Second, some sites or tissues in human body are immune privileged; they can tolerate self-immune threats. For example, tissue grafts in these sites are accepted easily. The sites/organs frequently mentioned in this context are brain, eye and reproductive organs, i.e., ovary and testis. Since brain and placenta are immune privileged and if CpG containing miRNAs could evoke some immune response like CpG containing oligo nucleotides, lack of immune response may relieve the selective pressure.

Although the mean values compared were constructed from random selections of CpG free microRNAs, the nature of Baskerville's data could create biased probabilities since the data contained a wide range of expression values per tissue. Therefore I normalized the data by taking natural logarithms of the intact values and of median divided values. Upon normalization, the probabilities approached to 0 except Brain and Frontal Cortex (Tables 4.4 and 4.5). Uterus, breast and placenta gave higher probabilities when compared to the results obtained using the raw data. These normalizations characterized the uterus, placenta, breast and lung as having low expression of CpG containing miRNAs, thus leaving brain and frontal cortex as the only tissues exhibiting no difference in expression of miRNAs with or without the CpG motif.

The results obtained from mouse were highly correlated with the findings in human. Except brain, in all mouse tissues tested there is a preference for CpG free microRNAs (Figure 4.14). In brain, there isn't any such preference. The commonalities between these two groups of results are remarkable. In both data, brain was different from other tissues, in the manner that preference on CpG-free

microRNA was less, and liver showed a striking discrepancy between the mean expression levels of CpG containing and CpG-free microRNAs.

Expression comparison of CpG versus non CpG microRNAs in staged mouse embryos and in mouse embryonic stem cells showed that mean expression values of both groups tended to increase over time. The probability of getting a higher mean expression from CpG free microRNAs is 0 after 10000 randomizations for mouse ES. This probability became higher, reaching to 0.611 and fell to 0.005 as the embryonic stages progressed.

From the immunity point of view, these results related to the embryonic data can be interpreted in the following way. The most vulnerable period for an embryo to autoimmune attacks, if probable, is the period before the placenta forms corresponding to the first stages of the embryonic development. If CpG containing microRNAs could evoke an immune response, the results mentioned above might be explained as an adaptation during microRNA evolution for escaping from self immune attacks. As the protection against the immune attacks forms, CpG containing microRNAs can be expressed freely.

Another finding was that the liver gave lower probabilities in all tests suggesting that presence of a CpG motif was highly negatively correlated with the expression level. The most frequently announced organ with autoimmune diseases is liver[8]. This finding may also supports the idea claimed in the previous paragraph. A wet lab work is required to test whether the CpG containing microRNAs evoke immune response or not. If microRNAs with CpG could evoke immune response, this could be one of the underlying mechanisms for many autoimmune diseases.

On the other hand, the testis gave significantly low probability indicating clear preference for CpG free miRNAs before and after normalizations, similar with other tissues responsible for reproduction. These findings seem to contradict the speculated idea of selective pressure against having CpG only in the non-immune

[8] Pubmed search: autoimmune diseases

protected tissues/organs. However, there may be other reasons that prevent the evolution of free utility of CpG containing microRNAs.

There is clearly CpG discrimination on the expression of microRNAs in most of the tissues. Immune response evoking potential of CpG dinucleotide is a possible speculation that can be made for explaining low expression levels of microRNAs. However, the methylation of cytosine in those CpG dinucleotides can be another mechanism that might down regulate the expression of miRNA species containing CpG motifs. Methylation can be a signal for the fate of microRNAs bearing it. Considering the effect of TpA on RNA stability, whether the expression levels of CpG containing microRNAs change with TpA existence also can be questioned.

Although the findings are clear, the data investigated are just constituted from two microarray studies. These results must further be confirmed using microarray data from other species, such as plant and invertebrate.

Localizations of CpG containing microRNAs should reveal whether there is a unique mechanism for microRNA transcriptional regulation. If a set of miRNAs are are clustered or within low expressed transcription units, then one can say that there a more coordinated transcriptional regulation exists.. In other cases, a requirement for individual transcriptional regulation could be suspected. Furthermore, the expression level comparisons of pre-mature microRNAs that contain CpG in their mature sequences should also be performed. Any difference between pre-mature and mature microRNA levels would indicate another mechanism apart from the transcriptional regulation. As a result, looking for the localizations of CpG containing microRNAs should be an immediate research and the present study may form a strong base for further studies.

REFERENCES:

Ahmad-Nejad, P., H. Hacker, M. Rutz, S. Bauer, R.M. Vabulas. and H. Wagner, 2002, Bacterial CpG-DNA and lipopolysaccharides activate Toll-like receptors at distinct cellular compartments, Eur. J. Immunol. 32, 1958.

Arhondakis, S., F. Auletta, G. Torelli and G. D'Onofrio, 2004, Base composition and expression level of human genes, Gene 325, 165.

Babak, T., W. Zhang, Q. Morris, B.J. Blencowe and T.R. Hughes, 2004, Probing microRNAs with microarrays: tissue specificity and functional inference, Rna 10, 1813.

Bartel, D.P., 2004, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116, 281.

Baskerville, S. and D.P. Bartel, 2005, Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes, Rna 11, 241.

Berezikov, E., V. Guryev, J. van de Belt, E. Wienholds, R.H. Plasterk and E. Cuppen, 2005, Phylogenetic shadowing and computational identification of human microRNA genes, Cell 120, 21.

Bernardi, G., 1993, The vertebrate genome: isochores and evolution, Mol Biol Evol 10, 186.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier, 1985, The mosaic genome of warm-blooded vertebrates, Science 228, 953.

Bourc'his D., G.L. Xu, C.S. Lin, B. Bollman and T.H. Bestor, 2001, Dnmt3L and the establishment of maternal genomic imprints. Science 294, 2536.

Calin, G.A., C.G. Liu, C. Sevignani, M. Ferracin, N. Felli, C.D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono, M.L. Dell'Aquila, H. Alder, L. Rassenti, T.J. Kipps, F. Bullrich, M. Negrini and C.M. Croce, 2004, MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias, Proc Natl Acad Sci U S A 101, 11755.

Castillo-Davis, C.I., S.L. Mekhedov, D.L. Hartl, E.V. Koonin and F.A. Kondrashov, 2002, Selection for short introns in highly expressed genes, Nat Genet 31, 415.

Cederberg, A., M. Hulander, P. Carlsson and S. Enerback, 1999, The kidney-expressed winged helix transcription factor FREAC-4 is regulated by Ets-1. A possible role in kidney development, J Biol Chem 274, 165.

Chedin F, M.R. Lieber and C.L. Hsieh, 2002, The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proc. Natl. Acad. Sci. USA 99 16916

Chen, C.Z., L. Li, H.F. Lodish and D.P. Bartel, 2004, MicroRNAs modulate hematopoietic lineage differentiation, Science 303, 83.

Dalpke, A. and K. Heeg, 2002, Signal integration following Toll-like receptor triggering, Crit Rev Immunol 22, 217.

Fitzpatrick D.R., Wilson C.B., 2003, Methylation and demethylation in the regulation of genes, cells and responses in the immune system. Clin. Immunol.109:37-45.

Galtier, N., G. Piganeau, D. Mouchiroud and L. Duret, 2001, GC-content evolution in mammalian genomes: the biased gene conversion hypothesis, Genetics 159, 907.

Gentles, A.J. and S. Karlin, 2001, Genome-scale compositional comparisons in eukaryotes, Genome Res 11, 540.

Gomez,R.A., M.L. Sequeira Lopez, L. Fernandez, D.R. Chernavvsky and V.F. Norwood, 1999, The maturing kidney: development and susceptibility. Ren. Fail. 21, 283.

Gonçalves, I., L. Duret, D. Mouchiroud, 2000, Nature and structure of human genes that generate retropseudogenes. Genome Res. 10, 672.

Hata K, M. Okano, H. Lei and E. Li, 2002, Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 129, 1983

Houbaviy, H.B., MZ.F. Murray and P.A. Sharp, 2003, Embryonic stem cell-specific MicroRNAs, Dev Cell 5, 351.

Jackson-Grusby, L., C. Beard, R. Possemato, M. Tudor, D. Fambrough, G. Csankovszki, J. Dausman, P. Lee, C. Wilson, E. Lander and R. Jaenisch, 2001, Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* 27, 31–39

Karlin, S. and J. Mrazek, 1997, Compositional differences within and between eukaryotic genomes, Proc Natl Acad Sci U S A 94, 10227.

Karlin, S., J. Mrazek and A.M. Campbell, 1997, Compositional biases of bacterial genomes and evolutionary implications, J Bacteriol 179, 3899.

Kim, J., A. Krichevsky, Y. Grad, G.D. Hayes, K.S. Kosik, G.M. Church and G. Ruvkun, 2004, Identification of many microRNAs that copurify with polyribosomes in mammalian neurons, Proc Natl Acad Sci U S A 101, 360.

Klinman, D.M., A.K. Yi, S.L. Beaucage, J. Conover and A.M. Krieg, 1996, CpG motifs present in bacteria DNA rapidly induce lymphocytes to secrete interleukin 6, interleukin 12, and interferon gamma, Proc Natl Acad Sci U S A 93, 2879.

Knox J. D., F.D. Araujo, P. Bigey, A.D. Slack, G.B. Price, M. Zannis-Hadjopoulos and M. Szyf, 2000 Inhibition of DNA methyltransferase inhibits DNA replication. *J. Biol. Chem.* 275 17986.

Konu, O. and M.D. Li, 2002, Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents, J Mol Evol 54, 35.

Krieg, A.M., 1996, An innate immune defense mechanism based on the recognition of CpG motifs in microbial DNA, J Lab Clin Med 128, 128.

Lagos-Quintana, M., R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel and T. Tuschl, 2002, Identification of tissue-specific microRNAs from mouse, Curr Biol 12, 735.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al., 2001, Initial sequencing and analysis of the human genome, Nature 409, 860.

Lau, N.C., L.P. Lim, E.G. Weinstein and D.P. Bartel, 2001, An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, Science 294, 858.

Lee P.P., D.R. Fitzpatrick, C. Beard, H.K. Jessup, S. Lehar, K.W. Makar, M. Perez-Melgosa, M.T. Sweetser, M.S. Schlissel, S. Nguyen, S.R. Cherry, J.H. Tsai, S.M. Tucker, W.M. Weaver, A. Kelso, R. Jaenisch and C.B. Wilson, 2001, A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. *Immunity.* 15 ,763.

Lee, R.C., R.L. Feinbaum and V. Ambros, 1993, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75, 843.

Lee Y, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, V.N. Kim, 2003, The Nuclear RNase III Drosha Initiates MicroRNA Processing, Nature 425, 415.

Lee, Y., K. Jeon, J.T. Lee, S. Kim and V.N. Kim, 2002, MicroRNA maturation: stepwise processing and subcellular localization, Embo J 21, 4663.

Li E., T.H. Bestor and R. Jaenisch, 1992, Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69 915

Liang, X.S., J.Q. Lian, Y.X. Zhou and M.B. Wan, 2004, Inhibitor RNA blocks the protein translation mediated by hepatitis C virus internal ribosome entry site in vivo, World J Gastroenterol 10, 664.

Lim, L.P., M.E. Glasner, S. Yekta, C.B. Burge and D.P. Bartel, 2003a, Vertebrate microRNA genes, Science 299, 1540.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge and D.P. Bartel, 2003b, The microRNAs of Caenorhabditis elegans, Genes Dev 17, 991.

Liu, C.G., G.A. Calin, B. Meloon, N. Gamliel, C. Sevignani, M. Ferracin, C.D. Dumitru, M. Shimizu, S. Zupo, M. Dono, H. Alder, F. Bullrich, M. Negrini and C.M. Croce, 2004, An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues, Proc Natl Acad Sci U S A 101, 9740.

Lund, E., S. Guttinger, A. Calado, J.E. Dahlberg and U. Kutay, 2004, Nuclear export of microRNA precursors, Science 303, 95.

Mouchiroud, D., G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier and G. Bernardi, 1991, The distribution of genes in the human genome, Gene 100, 181.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. 2000,. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408,** 86–89.

Pasquinelli, A.E. and G. Ruvkun, 2002, Control of developmental timing by micrornas and their targets, Annu Rev Cell Dev Biol 18, 495.

Ponger, L., L. Duret and D. Mouchiroud, 2001, Determinants of CpG islands: expression in early embryo and isochore structure, Genome Res 11, 1854.

Razzaque, M.S., T. Naito and T. Taguchi, 2001, Proto-oncogene Ets-1 and the kidney, Nephron 89, 1.

Robert M.F., S. Morin, N. Beaulieu, F. Gauthier, I.C. Chute, A. Barsalou and A.R. MacLeod, 2003, DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nature Genet. 33,. 61.

Rodriguez, A., S. Griffiths-Jones, J.L. Ashurst and A. Bradley, 2004, Identification of mammalian microRNA host genes and transcription units, Genome Res 14, 1902.

Saccone, S., C. Federico and G. Bernardi, 2002, Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds, Gene 300, 169.

Schmittgen, T.D., J. Jiang, Q. Liu and L. Yang, 2004, A high-throughput method to monitor the expression of microRNA precursors, Nucleic Acids Res 32, e43.

Sempere, L.F., S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky and V. Ambros, 2004, Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation, Genome Biol 5, R13.

Stacey, K.J., M.J. Sweet and D.A. Hume, 1996, Macrophages ingest and are activated by bacterial DNA, J Immunol 157, 2116.

Sugiyama, T., M. Gursel, F. Takeshita, C. Coban, J. Conover, T. Kaisho, S. Akira, D.M. Klinman and K.J. Ishii, 2005, CpG RNA: identification of novel single-stranded RNA that stimulates human CD14+CD11c+ monocytes, J Immunol 174, 2273.

Sun, Y., S. Koo, N. White, E. Peralta, C. Esau, N.M. Dean and R.J. Perera, 2004, Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs, Nucleic Acids Res 32, e188.

Thomson, J.M., J. Parker, C.M. Perou and S.M. Hammond, 2004, A custom microarray platform for analysis of microRNA gene expression, Nat Methods 1, 47.

Valoczi, A., C. Hornyik, N. Varga, J. Burgyan, S. Kauppinen and Z. Havelda, 2004, Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes, Nucleic Acids Res 32, e175.

Vinogradov, A.E., 2003, Isochores and tissue-specificity, Nucleic Acids Res 31, 5212.

Wightman, B., I. Ha and G. Ruvkun, 1993, Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans, Cell 75, 855.

Yi, A.K., D.M. Klinman, T.L. Martin, S. Matson and A.M. Krieg, 1996, Rapid immune activation by CpG motifs in bacterial DNA. Systemic induction of IL-6 transcription through an antioxidant-sensitive pathway, J Immunol 157, 5394.

Yi, R., Y. Qin, I.G. Macara and B.R. Cullen, 2003, Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs, Genes Dev 17, 3011.

Zhang, A., H. Jin, F. Zhang, Z. Ma, Y. Tu, Z. Ren, X. Zhang, K. Zhu and B. Wang, 2005, Effects of multiple copies of CpG on DNA vaccination, DNA Cell Biol 24, 292.

Zoubak, S., O. Clay and G. Bernardi, 1996, The gene distribution of the human genome, Gene 174, 95.

APPENDIX

mir.pl

```
#this PERL code calculates natural logarithms of dinucleotide and trinucleotide frequencies, multiplied
#by 100, for the pre-mature sequences (premr.txt) downloaded #from Sanger microRNA registry web site
#this code is also used for the calculation of the same values described above for mature sequences
#(mature.txt) and for the predicted sequences (prem.txt)
#the trinucleotidecalculations are skipped for the mature sequences

#opens premr,txt and records each row of the file as thee elements of array ‘line’
#and closes the file (the file is mature.txt in the case of mture sequences)
$mir='premr.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
print "$length\n";
$m=0;
close MIR;

#forms all possible dinucleotide motifs and records them as the elements of array #‘motifd’
@bases = split(/,/ , "A,T,G,C");
for ($p=0; $p<4; $p++) {
        for ($s=0; $s<4; $s++) {
    $motifd[$m]=$bases[$p].$bases[$s];
        $m=$m+1;
  }
}

#forms all possible trinucleotide motifs and records them as the elements of array #‘motift’

$m=0;
for ($p=0; $p<4; $p++) {
  for ($s=0; $s<4; $s++) {
    for ($t=0; $t<4; $t++) {
        $motift[$m]=$bases[$p].$bases[$s].$bases[$t];
        $m=$m+1;
  }
 }
}

# calculates natural logarithms of dinucleotide frequencies, multiplied by 100, of pre-#mature sequences for the 1st
frame and records them in txt file ‘dinuccleotidef1.txt’
$new='dinuccleotidef1.txt';
open (NEW,">$new");
        print NEW "UNIQID";
for ( $i=0; $i<16; $i=$i+1 ) {
        print NEW "\t$motifd[$i]";
}
for ($l=0; $l<$length; $l=$l+2) { #the increment is 4 for predicted sequences due to the nature of the file ‘prem.txt’
```

```
        @lin=split(/ /,$line[$l]);


        chomp $lin[0];
        print NEW "\n$lin[0]";
        for ( $i=0; $i<16; $i=$i+1 ) {
    $count=0;
        $c=0;
        $j=0;
        chomp $line[$l+1];
         $strlength=length $line[$l+1];
         while ($j<($strlength-2) ) {
      if (substr($line[$l+1],$j,2) eq $motifd[$i]) {
        $count=$count+1;
      }
                $c=$c+1;
       $j=$j+2;
    }
if ($count>0) {
                   $a=log(($count/$c)*100);
                }
                else {
            $a=0;
                }
            $rounded=sprintf("%.3f",$a);

                print NEW "\t$rounded";
        }
}
 close NEW;
```

calculates natural logarithms of dinucleotide frequencies, multiplied by 100, of pre-#mature sequences for the 2nd frame and records them in txt file ‘dinuccleotidef2.txt’

```
$new='dinuccleotidef2.txt';
open (NEW,">$new");
        print NEW "UNIQID";
for ( $i=0; $i<16; $i=$i+1 ) {
        print NEW "\t$motifd[$i]";
}
for ($l=0; $l<$length; $l=$l+2) {

        @lin=split(/ /,$line[$l]);

        chomp $lin[0];
        print NEW "\n$lin[0]";
        for ( $i=0; $i<16; $i=$i+1 ) {
    $count=0;
        $c=0;
        $j=1;
```

```
                chomp $line[$l+1];
                 $strlength=length $line[$l+1];
                 while ($j<($strlength-2) ) {
            if (substr($line[$l+1],$j,2) eq $motifd[$i]) {
               $count=$count+1;
            }
                        $c=$c+1;
             $j=$j+2;
        }
if ($count>0) {
                           $a=log(($count/$c)*100);
                        }
                        else {
                $a=0;
                        }
                $rounded=sprintf("%.3f",$a);
                        print NEW "\t$rounded";
            }
}
 close NEW;

#this part is removed for mature sequences
# calculates natural logarithms of trinucleotide frequencies, multiplied by 100, of pre-#mature sequences for the 1st frame and records them in txt file 'trinuccleotidef1.txt'

$new='trinucleotidef1.txt';
open (NEW,">$new");
            print NEW "UNIQID";
for ( $i=0; $i<64; $i=$i+1 ) {
            print NEW "\t$motift[$i]";
}
for ($l=0; $l<$length; $l=$l+2) {

            @lin=split(/ /,$line[$l]);
      chomp $lin[0];
            print NEW "\n $lin[0]";
            for ( $i=0; $i<64; $i=$i+1 ) {
         $count=0;
              $c=0;
              $j=0;
              chomp $line[$l+1];
              $strlength=length $line[$l+1];
               while ($j<($strlength-3)) {
                        if (substr($line[$l+1],$j,3) eq $motift[$i]) {
             $count=$count+1;
            }
            $j=$j+3;
            $c=$c+1;
        }
if ($count>0) {
                           $a=log(($count/$c)*100);
```

```
                        }
                        else {
                    $a=0;
                        }
                    $rounded=sprintf("%.3f",$a);
                        print NEW "\t$rounded";
            }
}
 close NEW;
```

calculates natural logarithms of trinucleotide frequencies, multiplied by 100, of pre-#mature sequences for the 2nd frame and records them in txt file ‘trinuccleotidef2.txt’

```
$new='trinucleotidef2.txt';
open (NEW,">$new");
            print NEW "UNIQID";
for ( $i=0; $i<64; $i=$i+1 ) {
            print NEW "\t$motift[$i]";
}
for ($l=0; $l<$length; $l=$l+2) {

            @lin=split(/ /,$line[$l]);
      chomp $lin[0];
            print NEW "\n $lin[0]";
            for ( $i=0; $i<64; $i=$i+1 ) {
        $count=0;
              $c=0;
              $j=1;
              chomp $line[$l+1];
              $strlength=length $line[$l+1];
               while ($j<($strlength-3)) {
                        if (substr($line[$l+1],$j,3) eq $motift[$i]) {
             $count=$count+1;

            }
            $j=$j+3;
            $c=$c+1;
         }
                         if ($count>0) {
                           $a=log(($count/$c)*100);
                        }
                        else {
                    $a=0;
                        }
                    $rounded=sprintf("%.3f",$a);
                        print NEW "\t$rounded";
            }
}
 close NEW;
```

calculates natural logarithms of trinucleotide frequencies, multiplied by 100, of pre-#mature sequences for the 3rd frame and records them in txt file ‘trinuccleotidef3.txt’

```
$new='trinucleotidef3.txt';
open (NEW,">$new");
        print NEW "UNIQID";
for ( $i=0; $i<64; $i=$i+1 ) {
        print NEW "\t$motift[$i]";
}
for ($l=0; $l<$length; $l=$l+2) {

        @lin=split(/ /,$line[$l]);

    chomp $lin[0];

        print NEW "\n $lin[0]";
        for ( $i=0; $i<64; $i=$i+1 ) {
      $count=0;
          $c=0;
          $j=2;
          chomp $line[$l+1];
           $strlength=length $line[$l+1];
           while ($j<($strlength-3)) {
                   if (substr($line[$l+1],$j,3) eq $motift[$i]) {
            $count=$count+1;

        }
        $j=$j+3;
        $c=$c+1;
       }
        if ($count>0) {
                   $a=log(($count/$c)*100);
                  }
                  else {
              $a=0;
                  }
              $rounded=sprintf("%.3f",$a);

                  print NEW "\t$rounded";
        }
}
 close NEW;
```

shuffle.pl

#this code shuffles the pre-mature sequences downloaded from Sanger miRNA #registry.(Also used for mature sequences by changing the input file as mature.txt)

```
#adds the shuffle function to the scope
use List::Util 'shuffle';

#opens mature.txt and records each row of the file as thee elements of array ‘line’
#and closes the file

$mir='premr.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
print "$length\n";
$m=0;
close MIR;

#creates a new file called shuffled.txt
$new='shuffled.txt';
open (NEW,">$new");
for ($l=0; $l<$length; $l=$l+2) {
    chomp $line[$l];
    chomp $line[$l+1];
    @lin=split(/ /,$line[$l]);
    @string=split(//,$line[$l+1]);
    print NEW "$lin[0]";

#shuffles the sequence in process
    @shuffled = shuffle(@string);
    $strings=join(//,@shuffled);
    $strings=~s/1//g;

#records the shuffled sequence to the shuffled.txt
print NEW "\n$strings\n";

}
close NEW;
```

shuffle.pl (modified for candidate genes)

```
#this code shuffles the candidate mature sequences predicted by Berezikov et al., (2005)

#adds the shuffle function to the scope
use List::Util 'shuffle';

#opens mature.txt and records each row of the file as thee elements of array ‘line’
#and closes the file
$mir='prem.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
print "$length\n";
close MIR;

#creates a new file called shuffled.txt
$new='shuffled.txt';
open (NEW,">$new");

for ($l=0; $l<$length; $l=$l+4) {

#renames candidate human genes
    $line[$l] =~ s/cand(\d\d )/cand0$1/;
    $line[$l] =~ s/cand(\d )/cand00$1/;

    chomp $line[$l];
    chomp $line[$l+1];
    @lin=split(/ /,$line[$l]);
    @string=split(//,$line[$l+1]);
    print NEW "$lin[0]";
    chomp $line[$l+1];

#shuffles the sequence in process
    @shuffled = shuffle(@string);
    $strings=join(//,@shuffled);
    $strings=~s/1//g;

#records the shuffled sequence to the shuffled.txt
print NEW "\n$strings\n";

}

close NEW;
```

mcount.pl

```
#this PERL code counts the mature microRNA sequences that don't contain certain #dinucleotide motifs

#opens mature.txt and records each row of the file as thee elements of array 'line'
#and closes the file
$mir='mature.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
close MIR;

#creates result file mcountr.txt
$new='mcountr.txt';
open (NEW,">$new");

@bases = split(/,/ , "A,T,G,C");
for ($p=0; $p<4; $p++) {
        for ($s=0; $s<4; $s++) {
    $motifd[$m]=$bases[$p].$bases[$s];
        $m=$m+1;
  }
}

#the first for loop changes the dinucleotide motif
for ( $k=0; $k<16; $k++ ) {
$mcount=0;

#the second for loop changes the sequence to be searched
for ($l=0; $l<$length; $l=$l+2) {

        @lin=split(/ /,$line[$l]);
        chomp $line[$l];
        chomp $line[$l+1];
        $b=(length $line[$l+1]);




          $j=0;
          $count=0;

          chomp $line[$l+1];
           $strlength=length $line[$l+1];
           while ($j<($strlength-1) ) {
        if (substr($line[$l+1],$j,2) eq $motifd[$k]) {
          $count=$count+1;
        }
                $j=$j+1;
    }
```

```
#if the motif looked for couldn’t be found than increase the counter
if ($count==0) {
$mcount++;
}
}
 print NEW "$mcount\t";
 }
close NEW;
```

first.pl

```
#this PERL code looks for each dinucleotide motifs whether they are at the 5'end of #the mature microRNA
#sequences

#opens mature.txt and records each row of the file as thee elements of array 'line'
#and closes the file
$mir='mature.txt';
open(MIR, $mir);
@line=<MIR>;
print "$line[1]";
$length=@line;
print "$length\n";
$m=0;
close MIR;

#creates result file 1st.txt
$new='1st.txt';
open (NEW,">$new");

#creates the dinucleotide motifs
@bases = split(/,/ , "A,T,G,C");
for ($p=0; $p<4; $p++) {
        for ($s=0; $s<4; $s++) {
    $motifd[$m]=$bases[$p].$bases[$s];
        $m=$m+1;
  }
}

$count=0;
for ( $i=0; $i<16; $i=$i+1 ) {
        #writes the header
        print NEW "\n$motifd[$i]\n";
$count=0;
for ($l=0; $l<$length; $l=$l+2) {

        @lin=split(/ /,$line[$l]);

        chomp $line[$l];
chomp $line[$l+1];
          if (substr($line[$l+1],0,2) eq $motifd[$i]) {
          $count=$count+1;
          #writes the members of the header
        print NEW " $line[$l]\n";
                }
}
print NEW "$count";
}
```

last.pl

```
#this PERL code looks for each dinucleotide motifs whether they are at the 3'end of #the mature microRNA
#sequences

#opens mature.txt and records each row of the file as thee elements of array 'line'
#and closes the file
$mir='mature.txt';
open(MIR, $mir);
@line=<MIR>;
print "$line[1]";
$length=@line;
print "$length\n";
$m=0;
close MIR;

#creates result file last.txt
open (NEW,">$new");
$new='last.txt';
@bases = split(/,/ , "A,T,G,C");
for ($p=0; $p<4; $p++) {
            for ($s=0; $s<4; $s++) {
      $motifd[$m]=$bases[$p].$bases[$s];
            $m=$m+1;
   }
}
$count=0;
for ( $i=0; $i<16; $i=$i+1 ) {

            #writes the header
            print NEW "\n$motifd[$i]\n";
$count=0;
for ($l=0; $l<$length; $l=$l+2) {

            @lin=split(/ /,$line[$l]);
            chomp $line[$l];
               chomp $line[$l+1];
               for ($z=0; $z<(length $line[$l+1]); $z=$z+1) {
                if (substr($line[$l+1],$z,4) eq $motifd[$i]) {
               $count=$count+1;

               #writes the members of the header
           print NEW " $line[$l]\n";

                        }
                        }
}
print NEW "$count";
}
```

vectors.pl

#this PERL code creates existence vectors for each of the four bases for four positions from #each ends; four #vectors #of 207x8

```
#opens mature.txt and records each row of the file as thee elements of array ‘line’
#and closes the file
$mir='mature.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
print "$length\n";
$m=0;
close MIR;

@bases = split(/,/ , "T,G,C,A");

#creates result file baevectors.txt
$new='baevectors.txt';
open (NEW,">$new");

$changes position
for ($a=0; $a<4; $a++) {
print NEW "Position $a\n";
print NEW "UNIQUE MIR ID\t5'T\tG\tC\tA\t3'T\tG\tC\tA\n";

        for ($l=0; $l<$length; $l=$l+2) {

                chomp $line[$l];
        chomp $line[$l+1];
        @lin=split(/ /,$line[$l]);
                print NEW "$line[$l]";
                $b=((length $line[$l+1])-1);



                for ($z=0; $z<4; $z++) {
                        if (substr($line[$l+1],$a,1) eq $bases[$z]) {
    print NEW "\t1";
  }

                else {
                        print NEW "\t0";

        }
                }
                for ($z=0; $z<4; $z++) {

                if (substr($line[$l+1],$b-$a,1) eq $bases[$z]) {
```

```
                print NEW "\t1";
            }
                        else {
                        print NEW "\t0";

                        }
                        }
print NEW "\n";
}
print NEW "\n";
}
close NEW;
```

jaccard.m

```
% This Matlab 6.22 code calculates Jaccard similarity between columns of a given %vector x

function z = jakard(x)
[a,b]=size(x);
for c=1:b
   countone=0;
   counttwo=0;
   countco=0;
   for r=1:a
      if x(r,c)==1
         countone=countone+1;
      end
   end
      for d=1:b
         counttwo=0;
         countco=0;
         for r=1:a
            if x(r,d)==1
               counttwo=counttwo+1;
            end
            if x(r,c)==1 & x(r,d)==1
               countco=countco+1;
            end
         end
         j=countco/(countone+counttwo-countco);
         z(d,c)=j;
      end
   end
```

cmirs.pl

```
#this PERL code derives the common micoRNAs between registry and expression data by Baskerville et al.,
#2005 #as names and records them to common.txt

#opens the new file, common.txt
$new='common.txt';
open (NEW,">$new")
;
$mire='h2.txt'; #the file that contain the names of expressed microRNAs
open(MIR, $mire);
@exprressed=<MIR>;

$nmir='nmature.txt'; #the file that contain the names of registered microRNAs
open(NMIR, $nmir);
@nmature=<NMIR>;
close MIR;
close NMIR;

$lone=@nmature;
$ltwo=@exprressed;

for ($i=0; $i<$ltwo; $i++) {
        for ($j=0; $j<$lone; $j=$j+2) {
                if ($exprressed[$i] eq $nmature[$j]) {
                  print NEW $exprressed[$i];
                }
        }
}
close NEW;
```

cmirs2.pl

#this PERL code extracts the common microRNAs between registry and expression data by Baskerville et. al., 2005 #and writes them to a new file with their expression #values

```
$new='common2.txt';
open (NEW,">$new");
$mire='human.txt';
open(MIR, $mire);
@exprressed=<MIR>;
$nmir='nmature.txt';
open(NMIR, $nmir);
@nmature=<NMIR>;
close MIR;
close NMIR;

$lone=@nmature;
$ltwo=@exprressed;

for ($i=0; $i<$ltwo; $i++) {
        for ($j=0; $j<$lone; $j=$j+2) {
                @str=split(/\t/ , $exprressed[$i]);
                chomp $nmature[$j];
                $nmature[$j]=~s/\>//g;
                if ( $str[0] eq $nmature[$j]) {
                  print NEW $exprressed[$i];
                        print "alal"
                }
        }
}
close NEW;
```

cpgn.pl

```
#this PERL code separates CpG containing and CpG free common microRNAs

$new='cpgc.txt';
open (NEW,">$new");
$nnew='ncpgc.txt';
open (NNEW,">$nnew");
$mire='common.txt';
open(MIR, $mire);
@exprressed=<MIR>;
$nmir='nmature.txt';
open(NMIR, $nmir);
@nmature=<NMIR>;
close MIR;
close NMIR;

$lone=@nmature;
$ltwo=@exprressed;

for ($i=0; $i<$ltwo; $i++) {
        for ($j=0; $j<$lone; $j=$j+2) {
                if ($exprressed[$i] eq $nmature[$j]) {
                        $mirl=length $nmature[$j+1];
            while ((substr($nmature[$j+1],$k,2) ne 'CG')&&($k<($mirl-1))) {
                $k=$k+1;

      }
                @names=split(/\t/,$nmature[$j]);
                if ($k<($mirl-1)) {
                print NEW "$names[0]";
                }
                else {
                print NNEW "$names[0]";
                }
                $k=0;

                }
        }
}
close NEW;
```

cpg.pl (used for all data)

```
#this PERL code separates CpG containing and CpG free common microRNAs
#with their expression values

$new='cpg.txt';
open (NEW,">$new");
$nnew='ncpg.txt';
open (NNEW,">$nnew");
$mire='common2.txt';
open(MIR, $mire);
@exprressed=<MIR>;
$nmir='nmature.txt';
open(NMIR, $nmir);
@nmature=<NMIR>;
close MIR;
close NMIR;

$lone=@nmature;
$ltwo=@exprressed;

for ($i=0; $i<$ltwo; $i++) {
        for ($j=0; $j<$lone; $j=$j+2) {
                chomp $nmature[$j];
                $nmature[$j]=~s/\>//g;
                @emir=split(/\t/ , $exprressed[$i]);
                if ($emir[0] eq $nmature[$j]) {
                        $mirl=length $nmature[$j+1];
           while ((substr($nmature[$j+1],$k,2) ne 'CG')&&($k<($mirl-1))) {
                $k=$k+1;

      }

                if ($k<($mirl-1)) {
                print NEW "$exprressed[$i]";
                }
                else {
                print NNEW "$exprressed[$i]";
                }
                $k=0;

                }
        }
}
close NEW;
```

cpgmean.pl (used for all data)

```
#calculates the mean expression values of CpG containing and CpG free microRNAs for each tissue

$hum='human.txt';
open(HUM, $hum);
@columns=<HUM>;
$lengthc=@columns;
chomp $columns[0];
@tissues=split(/\t/,$columns[0]);
$lengtht=@tissues;
shift @columns;

$new='ncpgmeansa.txt';
open(NEW,">$new");

$mir='ncpgc.txt';
open(MIR, $mir);
@cpg=<MIR>;
$lengthcpg=@cpg;

for ($a=1; $a < ($lengtht); $a++) {

for ($i=0; $i<$lengthcpg; $i++) {
chomp $cpg[$i];
$cpg[$i]=~s/\>//g;

$scanner=0;

@values=split(/\t/,$columns[$scanner]);
while ($values[0] ne $cpg[$i]) {
$scanner=$scanner+1;
@values=split(/\t/,$columns[$scanner]);
}

chomp $values[$a];

${$tissues[$a]}=${$tissues[$a]} + $values[$a];
}
$ali=${$tissues[$a]}/$lengthcpg;
print NEW $ali, "\t";
${$tissues[$a]}=0;
}
```

randomly.pl (used for all data except for the one to be logged)

#selects a number of CpG free microRNAs equals to number of CpG containing ones and calculates the expression #mean of the selected mirs

```
#adds the shuffle function to the scope
use List::Util 'shuffle';

$hum='human.txt';
open(HUM, $hum);
@columns=<HUM>;
$lengthc=@columns;
chomp $columns[0];
@tissues=split(/\t/,$columns[0]);
$lengtht=@tissues;
shift @columns;

$new='means.txt';
open (NEW,">$new");

$mir='ncpgc.txt';
open(MIR, $mir);
@ncpg=<MIR>;

for ($i=0; $i<98; $i++) {  #98 is the number of non-CpG microRNAs and it changes #according to data
$array[$i]=$i;
}

for ($r=0; $r<1000; $r++) { #randomization number
print $r, "\n";
@shuffled = shuffle(@array);

for ($i=0; $i<50; $i++) { # 50 is the number of CpG containing microRNAs and it #changes according to data
$scanner=0;

@values=split(/\t/,$columns[$scanner]);
chomp $ncpg[$shuffled[$i]];
while (($values[0] ne $ncpg[$shuffled[$i]]) ) {
$scanner=$scanner+1;
@values=split(/\t/,$columns[$scanner]);
}

for ($a=1; $a < 26; $a++) {
$scanner=0;
chomp $values[($a)];
# $values[($a)] is logged when the log normalization is necessary
${$tissues[$a]}=${$tissues[$a]} + $values[($a)];

if ($values[$a] ne 'NaN') {
$counter{$a}=$counter{$a}+1;
```

```
}
}
}
for ($a=1; $a < 26; $a++) {

$ali=${$tissues[$a]}/$counter{$a};
print NEW $ali, "\t";
${$tissues[$a]}=0;
$counter{$a}=0;

}
print NEW "\n";

}
```

probable.pl

```
#calculates the probability of CpG containing means being higher than the means of randomly selected CpG free #microRNAs

$meane='means.txt';
open(MEA, $meane);
@columns=<MEA>;

$new='p.txt';
open (NEW,">$new");

$mir='cpgmeans.txt';
open(MIR, $mir);
@cpgc=<MIR>;
@cpg=split(/\t/,$cpgc[0]);
for ($a=0; $a < 25; $a++) {
$counter{$a}=0;

for ($i=0; $i<1000; $i++) {

@values=split(/\t/,$columns[$i]);
if ($values[$a] < $cpg[$a]) {
$counter{$a}=$counter{$a}+1;
}
}

$ali=$counter{$a}/1000;
print NEW $ali, "\t";
}
```

median.pl

```
#takes the median values of the columns of the expression data

sub mean {
	my(@data)=@_;
	my $sum;
	foreach(@data) {
		$sum+=$_;
	}
	return($sum/@data);
}

sub median {
	my(@data)=sort { $a <=> $b} @_;
	if (scalar(@data)%2) {
		return($data[@data/2]);
	} else {
		my($upper, $lower);
		$lower=$data[@data/2];
		$upper=$data[@data/2 - 1];
		return(mean($lower, $upper));
	}
}

$hum='human.txt';
open(HUM, $hum);
@columns=<HUM>;
$lengthc=@columns;
chomp $columns[0];
@tissues=split(/\t/,$columns[0]);
$lengtht=@tissues;
print $lengtht;
shift @columns;

$new='medians.txt';
open (NEW,">$new");

for ($a=1; $a < ($lengtht); $a++) {
for ($i=0; $i<($lengthc-1); $i++) {
@values=split(/\t/,$columns[$i+1]);
chomp $values[$a];

$array[$i]=$values[$a];
}
print NEW median(@array), "\t";
}
```

medians.pl

```
#converts the expression data to median substracted log normalized one
#records it to humans.txt

$hum='human.txt';
open(HUM, $hum);
@columns=<HUM>;
$lengthc=@columns;
@tissues=split(/\t/,$columns[0]);
$lengtht=@tissues;
$new='humans.txt';
open (NEW,">$new");
print NEW $columns[0];
shift @columns;


$mir='medians.txt';
open(MIR, $mir);
@medians=<MIR>;
@median=split(/\t/,$medians[0]);
for ($i=0; $i<$lengthc; $i++) {
@values=split(/\t/,$columns[$i]);
print NEW "$values[0]\t";
for ($a=0; $a < ($lengtht-1); $a++) {
chomp $values[$a+1];
chomp $median[$a];
if (($values[$a+1]/$median[$a])==0) {
print NEW "NaN\t";
}
else {
$division=log($values[$a+1]/$median[$a]);
print NEW "$division\t";
}
}
print NEW "\n";
}
close NEW;
```

newshuffle.pl

 # this code shuffles the sequences calculates the means of the logs of the 100 #multiplied dinucleotide frequencies and repeats this procedure at randomization #number times

```
#adds the shuffle function to the scope
use List::Util 'shuffle';

$mir='mature.txt';
open(MIR, $mir);
@line=<MIR>;
$length=@line;
close MIR;
$new='thematrix.txt';
open (NEW,">$new");

@bases = split(/,/ , "A,T,G,C");
for ($p=0; $p<4; $p++) {
        for ($s=0; $s<4; $s++) {
    $motifd[$m]=$bases[$p].$bases[$s];
        $m=$m+1;
  }
}
for ($random=0; $random<1000; $random++) {
for ($l=0; $l<$length; $l=$l+2) {
                @lin=split(/ /,$line[$l]);
                chomp $line[$l];
        chomp $line[$l+1];
        $b=(length $line[$l+1]);

for ($i=0; $i<$b; $i++) {
$arr[$i]=$i;
}
@shuffled=shuffle(@arr);
@str=split(//, $line[$l+1]);

for ($i=0; $i<$b; $i++) {
$string.=$str[$shuffled[$i]];
}

for ( $k=0; $k<16; $k++ ) {
      $count=0;
          $c=0;
          $j=0;
          chomp $line[$l+1];
           $strlength=length $line[$l+1];
           while ($j<($strlength-2) ) {
        if (substr($string,$j,2) eq $motifd[$k]) {
          $count=$count+1;
        }
```

```
                $c=$c+1;
            $j=$j+1;
        }
if ($count>0) {
                $a=log(($count/$c)*100);
                }
                else {
            $a=0;
                }
            $rounded=sprintf("%.3f",$a);

${$motifd[$k]}[$l/2]=$rounded;
            }
$string="";
}
for ( $k=0; $k<16; $k++ ) {

 foreach (@{$motifd[$k]}) {
$sum+=$_;
}
@{$motifd[$k]}="";
$z=$sum/($l/2);
print NEW "$z\t";
$sum=0;
}
 print NEW "\n";
}
close NEW;
```

pvalue.pl

```
#calculates how many higher `or lower (if given number lower or higher respectively)  value exists in each
#column of a matrix than the values in the corresponding columns of a given matrix and calculates the
#probability

$actuals='a.txt'; #given one-row matrix
open(A, $actuals);
$a=<A>;
close A     ;
@avalues=split(/\t/, $a);
$i=0;
$montecarlo='thematrix.txt'; # many rows, output of newshuffle.pl
open(B, $montecarlo);
while ($b=<B>) {
@{$i}=split(/\t/, $b);
$i++;
}
close B;
$new='p.txt';
open (NEW,">$new");

for ($c=0; $c<16; $c++) {
        for ($e=0; $e<1001; $e++) {
                $bvalues{$c}[$e]=${$e}[$c];
        }
}

for ($c=0; $c<16; $c++) {
        if ($avalues[$c]<$bvalues{$c}[1000]) {
                for ($e=0; $e<1000; $e++) {
                        if ($avalues[$c]>=$bvalues{$c}[$e]) {
                                $count{$c}++;
                        }
                }
        }
        elsif ($avalues[$c]>$bvalues{$c}[1000]) {
                for ($e=0; $e<1000; $e++) {
                        if ($avalues[$c]<=$bvalues{$c}[$e]) {
                                $count{$c}++;
                        }
                }
        }
$p=$count{$c}/1000;
print NEW "$p \t";
}
close NEW;
```