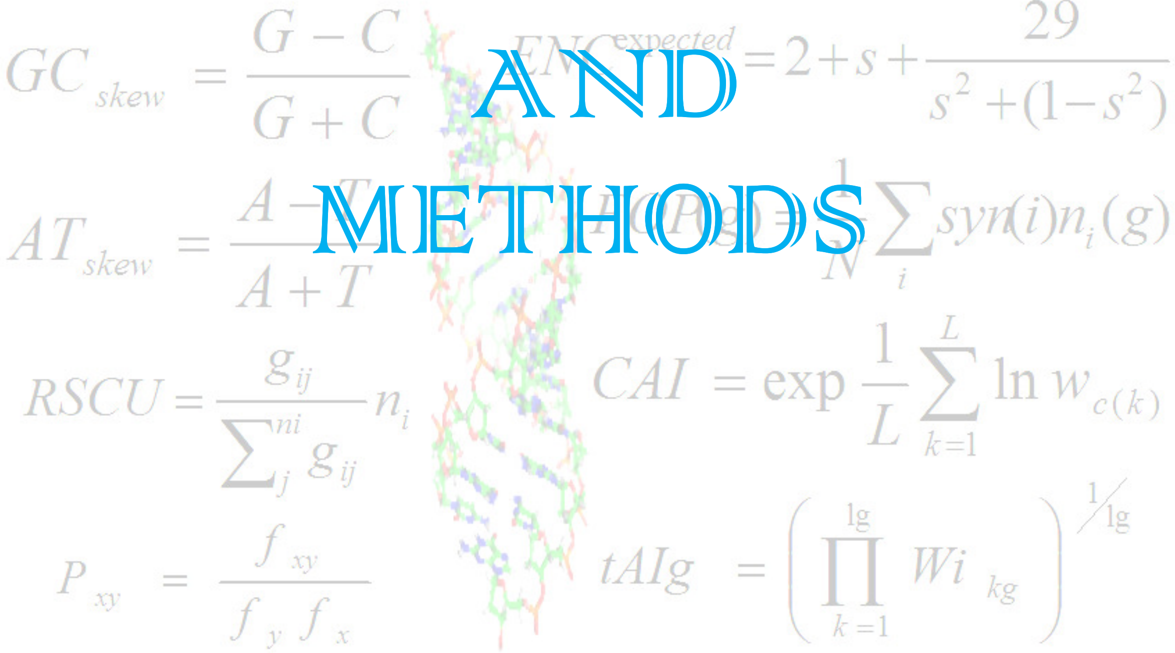
CHAPTER – 3

MATERIALS AND METHODS


$GC_{skew} = \frac{G-C}{G+C}$
$AT_{skew} = \frac{A-T}{A+T}$
$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} n_i$
$P_{xy} = \frac{f_{xy}}{f_y f_x}$
$ENC^{expected} = 2+s+\frac{29}{s^2+(1-s^2)}$
$(Pg) = \frac{1}{N}\sum_i syn(i)n_i(g)$
$CAI = \exp\frac{1}{L}\sum_{k=1}^{L}\ln w_{c(k)}$
$tAIg = \left(\prod_{k=1}^{lg} Wi_{kg}\right)^{1/lg}$

3. Materials and Methods

3.1. Sequence Data and Resources

Human proto-oncogene, oncogene and tumour suppressor gene record was obtained from the web site (http://cbio.mskcc.org/CancerGenes) (Hahn and Weinberg 2002, Higgins *et al*. 2007). A literature search was performed to determine the role of the specific gene in cancer development. Any genes whose role as a proto-oncogene/oncogene or tumour suppressor was still questionable were not included in the final list. Complete nucleotide coding sequence (cds) for each gene was then retrieved from National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov). We selected only those coding sequences which are an exact multiple of 3-base with perfect start codon (ATG, GTG, CTG or TTG) at the beginning of cds and stop codon (TAA, TAG or TGA) at the end of cds, devoid of any unknown bases (N) and any intercalary stop codon in the entire sequence. Finally, we selected eighty-two coding sequences for proto-oncogenes/oncogenes and sixty-three coding sequences for tumour suppressor genes that fulfill the aforementioned criteria and used in our CUB analysis (Table 1, Table 2).

Table 1: Accession numbers of coding sequences (cds) of eighty-two human proto-oncogenes/oncogenes with their lengths (bp)

CDS NO.	GENES	ACCESSION NO.	LENGTH (bp)
1	*ABL (ABL1)*	M14752.1	3393
2	*ABL2*	DQ009672.1	3549
3	*AKT1*	AH011307.1	1443
4	*AKT2*	BC063421.1	444
5	*ATF1*	BC029619.1	816
6	*BCL11A*	GU324937.1	2508
7	*BCL2*	AY220759.1	720
8	*BCL3*	M31732.1	1341
9	*BCL6*	EU883531.1	1953
10	*BCR*	U07000.1	3816
11	*BRAF*	EU600171.1	2301
12	*CARD11*	BC111719.1	3444
13	*CBLB*	U26710.1	2949
14	*CBLC*	BC006122.1	678
15	*CCND1* (*Cyclin D1*)	M64349.1	888
16	*CCND2* (*Cyclin D2*)	M90813.1	870
17	*CCND3* (*Cyclin D3*)	M90814.1	879
18	*CDX2*	KJ081251.1	816
19	*CTNNB1*	AY463360.1	2346
20	*DDB2*	AY220533.1	1284
21	*DDIT3*	AY880949.1	510
22	*DDX6*	BC039826.1	564
23	*DEK*	BC035259.1	1128
24	*EGFR*	U48722.1	1218
25	*ELK4*	BC063676.1	1218
26	*ERBB2*	AY208911.1	3768
27	*ETV4*	BC016623.1	1455
28	*ETV6*	BC043399.1	1359
29	*EVI1*	GQ352634.1	3156
30	*EWSR1*	BC011048.1	1968
31	*FEV*	BC023511.2	717
32	*FGFR1*	AY585209.1	2469
33	*FGFR1OP*	BC037785.1	450
34	*FGFR2*	M97193.1	2469
35	*FUS*	CR456747.1	1581
36	*GOLGA5*	BC023021.1	2196
37	*GOPC*	KF420123.1	1224
38	*HMGA1*	BC067083.1	324
39	*HMGA2*	AY601863.1	279
40	*HRAS*	EF015887.1	513

Table 1 continued

CDS NO.	GENES	ACCESSION NO.	LENGTH (bp)
41	*IRF4*	BC015752.1	1356
42	*JUN*	J04111.1	996
43	*KIT*	U63834.1	2931
44	*KRAS*	JX512447.1	570
45	*LCK*	M36881.1	1530
46	*LMO2*	BC034041.1	477
47	*MAF*	AF540388.1	1113
48	*MAFB*	BC036689.1	972
49	*MAML2*	AY040322.1	3462
50	*MDM2*	GQ848196.1	1401
51	*MET*	J02958.1	4227
52	*MITF*	BC065243.1	1260
53	*MLL*	BT007215.1	1707
54	*MPL*	M90103.1	1740
55	*MYB*	AF104863.1	1923
56	*MYC*	AY214166.1	1320
57	*MYCL1*	BC011864.2	621
58	*MYCN*	BC002712.2	1395
59	*NCOA4*	BC001562.1	1845
60	*NFKB2*	AY865619.1	2703
61	*NRAS*	EU332857.1	570
62	*NTRK1*	BC136554.1	2373
63	*NUP214*	BC105998.1	6243
64	*PAX8*	L19606.1	1353
65	*PDGFB*	BC077725.1	726
66	*PIK3CA*	BC113601.1	3207
67	*PIM1*	M27903.1	942
68	*PLAG1*	BC075047.2	1503
69	*PPARG*	AB451337.1	1518
70	*PTPN11*	BC030949.1	153
71	*RAF1*	AY271661.1	1947
72	*REL*	DQ314888.1	1860
73	*RET*	BC004257.1	3219
74	*ROS1*	M34353.1	7044
75	*SMO*	AH007453.1	2364
76	*SS18*	BC096222.1	1257
77	*TCL1A*	BC014024.1	345
78	*TET2*	BC150180.1	1914
79	*TFG*	BC041600.1	1101
80	*TLX1*	BC130530.1	774
81	*TPR*	U69668.1	7092
82	*USP6*	AY143550.1	4221

CDS NO: Coding sequence number; Yellow color marked genes represent proto-oncogenes

Table 2: Accession numbers of coding sequences (cds) of sixty-three human tumour suppressor genes with their lengths (bp)

CDS NO.	GENES	ACCESSION NO.	LENGTH (bp)
1	*APC*	M74088.1	8532
2	*ARHGEF12*	NG_027960.1	4635
3	*ATM*	U33841.1	9171
4	*BCL11B*	NM_138576.3	2685
5	*BLM*	AY886902.1	4254
6	*BMPR1A*	BC028383.1	1599
7	*BRCA1*	U14680.1	5592
8	*BRCA2*	DQ897648.1	7950
9	*CARS*	BC002880.2	2247
10	*CBFA2T3*	BC062624.1	1848
11	*CDH1*	GU371438.1	2649
12	*CDH11*	BC013609.1	2391
13	*CDK6*	AY128534.1	981
14	*CDKN2C*	AY094608.1	507
15	*CEBPA*	EU048234.1	1077
16	*CHEK2*	CU012979.1	1761
17	*CREB1*	BC095407.1	984
18	*CREBBP (CBP)*	U47741.1	7329
19	*CYLD*	BC012342.1	2862
20	*DDX5*	BC016027.1	1845
21	*EXT1*	BC001174.1	2241
22	*EXT2*	U62740.1	2157
23	*FBXW7*	BC117246.1	2124
24	*FH*	U59309.1	1533
25	*FLT3*	BC126350.1	2982
26	*FOXP1*	BC071893.1	345
27	*GPC3*	L47125.1	1743
28	*IDH1*	BC093020.1	1245
29	*IL2*	K02056.1	462
30	*JAK2*	AY973034.1	3399
31	*MAP2K4*	DQ015703.1	1200
32	*MDM4*	BC067299.1	1473
33	*MEN1*	U93237.1	1848
34	*MLH1*	BC006850.1	2271
35	*MSH2*	AY601851.1	2805
36	*NF1*	M89914.1	8520
37	*NF2*	BC020257.1	1788
38	*NOTCH1*	CR457221.1	426
39	*NPM1*	M28699.1	885
40	*NR4A3*	NG_028910.1	1914

Table 2 continued

CDS NO.	**GENES**	**ACCESSION NO.**	**LENGTH (bp)**
41	*NUP98*	U41815.1	2763
42	*PALB2 (BRCA2)*	BC044254.2	3561
43	*PML*	BC000080.2	2346
44	*PTEN*	DQ073384.1	1212
45	*RB1*	BC039060.1	2787
46	*RUNX1*	BC136381.1	1443
47	*SDHB*	U17248.1	843
48	*SDHD*	BC015992.1	480
49	*SMARCA4*	BC136644.1	5046
50	*SMARCB1*	BC143667.1	1131
51	*SOCS1*	DQ086801.1	636
52	*STK11 (LKB1)*	BC019334.1	1302
53	*SUFU*	BC013291.2	1455
54	*SUZ12*	NM_015355.2	2220
55	*SYK*	BC001645.2	1908
56	*TCF3*	BC110580.1	1977
57	*TNFAIP3*	BC114480.1	2373
58	*TP53*	U94788.1	1182
59	*TSC1*	AB190910.1	1365
60	*TSC2*	BC150300.1	5355
61	*VHL*	AF010238.1	642
62	*WRN*	AY442327.1	4299
63	*WT1*	AY245105.1	1350

CDS NO: Coding sequence number

3.2. Analysis of synonymous codon usage bias

Two amino acids methionine and tryptophan are encoded by single codon ATG and TGG, respectively and three stop codons (TAA, TAG, and TGA) would not reveal any usage bias and therefore discarded from the calculation. We measured the non-uniform usage of synonymous codons for the proto-oncogenes, oncogenes and tumour suppressor genes by analyzing several genetic indices given below-

3.2.1. *Nucleotide Composition Analysis*

Compositional properties for each of the selected cds were calculated as follows:

(a) Occurrence of the nucleotides (A, T, G, C) overall frequency.

(b) Frequency of the nucleotide at third position of synonymous codon (A_3, T_3, G_3, C_3).

(c) Overall GC and AT contents.

(d) Occurrence of overall frequency of the nucleotide (G+C) % at first (GC_1%), second (GC_2%) and third (GC_3%) position of synonymous codon.

(e) Average of the nucleotide G+C contents at first and second (GC_{12}) position of synonymous codons.

3.2.2. *Analysis of nucleotide skewness*

Nucleotide composition in the leading and lagging strands of DNA usually differs as a result of asymmetry in biochemical processes such as DNA replication and repair (Sueoka 1962). This creates a variation in synonymous codon usage between leading and lagging strands. The distribution of different oligomers is skewed between leading and lagging DNA strands (Salzberg *et al*. 1998), but the trend is most obvious when considering a difference in the number of Gs and Cs along a single DNA strand. This difference usually measured as:

$$GC_{skew} = \frac{G-C}{G+C}$$

Positive GC skew indicates richness of G over C and negative as C over G (Tillier and Collins 2000). Similarly, AT skew also varies between leading and lagging strand and can be defined as:

$$AT_{skew} = \frac{A-T}{A+T}$$

Positive AT skew represents overloading of A over T and negative with overloading T over A (Tillier and Collins 2000).

3.2.3. *Effective Number of Codons (ENC) Analysis*

ENC is generally used to quantify the codon usage bias of a gene that is independent of the gene length and number of amino acids (Wright 1990). The values of *ENC* ranged from 20 indicating strong codon bias in the gene using only one synonymous codon within a family for the corresponding amino acid, to 61 indicating no bias in the gene using all synonymous codons equally for the corresponding amino acid (Wright 1990). This measure was computed as per Wright (1990) to estimate the codon usage affected by GC_{3s} under mutation pressure or genetic drifts, among the coding sequences of proto-oncogene/oncogenes and tumour suppressor genes:

$$ENC^{\exp ected} = 2 + s + \frac{29}{s^2 + (1-s^2)}$$

where *s* denotes the given GC_3% values (Wright 1990).

3.2.4. *Frequency of Optimal Codon (Fop) Analysis*

Fop is a measure of codon usage bias in a gene (Ikemura 1985). *Fop* values represent the ratio of the number of optimal codons used to the total number of synonymous codons (Ikemura 1981). The *Fop* value ranges from 0.36 for a gene showing uniform codon usage bias to 1 for a gene showing strong codon usage bias (Stenico *et al*. 1994). *Fop* value for each selected coding sequence was calculated using the formula given by Lavner and Kotler (2005) which is as follows:

$$FOP_s(g) = \frac{1}{N}\sum_i n_i(g)$$

where the subscript 's' stands for "simple" and $n_i(g)$ is the count of the codon i in the gene g , N is the total number of codons and sum is taken over all the optimal codons (Lavner and Kotlar 2005). The *FOP* measure in this way is affected by amino acid usage because if synonymous codon usage is random, the 2-fold degeneracy of amino acids would have *FOP* value 0.5, whereas 4-fold degeneracy of amino acids would have *FOP* value 0.25.

Therefore, in order to obtain a measure which is independent of amino acid composition, each codon count in the above equation is multiplied by the corresponding amino acid.

$$FOP(g) = \frac{1}{N}\sum_i syn(i)n_i(g)$$

where $syn(i)$ is the degeneracy of the amino acid coded by i . Thus, in this way a gene close to random synonymous codon usage will have *FOP* value close to 1 regardless of its amino acid composition (Lavner and Kotlar 2005).

3.2.5. *Relative Synonymous Codon Usage (RSCU) Analysis*

RSCU is defined as the observed frequency of a codon divided by the expected frequency if all codons are used equally for any particular amino acid (Sharp and Li 1986a). *RSCU* values of codons for all the selected cds of tumour suppressor genes were calculated as follows:

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} n_i$$

where g_{ij} is the observed number of the *i*th codon for the *j*th amino acid which has n_i kinds of synonymous codons (Butt *et al.* 2014).

The *RSCU* value equal to unity represents that the codons are used equally or randomly for the corresponding amino acid and no bias for that amino acid (Sharp and Li 1986b). The *RSCU* value greater than one represents positive codon usage bias with greater usage of the most abundant codons whereas the *RSCU* value less than one represents a negative codon usage bias using the least abundant codons.

Moreover, synonymous codons with *RSCU* values >1.6 are considered as over-represented codons and <0.6 as under-represented codons, respectively (Wong *et al.* 2010).

3.2.6. *Relative dinucleotide abundance*

The relative abundance of dinucleotide in the coding sequences of proto-oncogene/oncogene and tumour suppressor gene in human was calculated using the approach of Chiusano and his co-workers (Chiusano et al. 2000). The odd ratio for each dinucleotide was computed using the following formula:

$$P_{xy} = \frac{f_{xy}}{f_y f_x}$$

where f_x and f_y denotes the frequency of the nucleotide X and Y respectively, and f_{xy} denotes the frequency of the dinucleotide XY. In our analysis, the dinucleotide with p_{xy} >1.23 is considered to be over-represented dinucleotide whereas p_{xy} <0.78 as under-represented dinucleotide in terms of relative abundance.

3.2.7. *Correspondence Analysis*

Correspondence analysis is generally used to investigate the major trend in codon usage variation among genes and distributes the codons in axis1 and axis2 with these trends (Perriere and Thioulouse 2002, Wang and Hickey 2007). To explore the variation in codon usage among human proto-oncogene/oncogenes and tumour suppressor genes, *RSCU* values of all cds selected in this study were used for correspondence analysis. Each cds was represented as a 59-dimensional vector, and each dimension corresponds to the *RSCU* value of one sense codon with the exception of ATG (methionine), TGG (tryptophan) and three stop codons.

3.2.8. *Codon Adaptation Index (CAI) Analysis*

CAI is a quantitative measure used to predict the level of gene expression on the basis of extent of bias towards codon sequence. The *CAI* values ranged from 0 for a gene (exhibiting no bias using all possible synonymous codons equally for the corresponding amino acid) to 1 for a gene (exhibiting strong codon bias using only one possible synonymous codon for the corresponding amino acid). The *CAI* value was measured as per Sharp and Li (1987) which was as follows:

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_{c(k)}$$

where L is the number of codons in the gene and $w_{c(\mathrm{k})}$ is the ω (relative adaptiveness) value for the *k*-th codon in the gene (Sharp and Li 1987).

In our study *CAI* value was calculated using the approach of Puigbo *et al.*, for human codon usage as reference set (Puigbo *et al.* 2008) (available at: http://genomes.urv.es/CAIcal).

3.2.9. *tRNA Adaptation Index (tAI)*

The parameter *tAI* measures the degree of adaptation of a gene (cds) to its genomic *tRNA* pool available for translation. It estimates the accessibility of *tRNAs* for each codon. Since an anticodon can identify several codons with different competence weights due to wobble interactions and hence the codon-anticodon pairing is not unique in terms of base complementarities in the coding sequence of the gene. Therefore, *tAI* is a good measure for predicting the translational selection in a genome which was formulated by dos Reis *et al.*, as follows:

$$tAIg = \left(\prod_{k=1}^{\mathrm{lg}} Wi_{kg} \right)^{1/\mathrm{lg}}$$

where Wi_{kg} is the relative adaptiveness value of the codon defined by the k_{th} position of the gene *g* and *lg* is the length of the gene *g* in terms of codon (excluding stop codons) (dos Reis et al. 2004). The *tAI* value for each cds of the selected proto-oncogene/oncogene and tumour suppressor gene pool was calculated using human *tRNA* gene copy numbers and implemented in visual gene developer 1.7 software (Jung and McDonald 2011).

3.3. Analysis

3.3.1. Correlation analysis

Correlation analysis was used to identify the relationship between the pattern of synonymous codon usage and the genetic indices used for the present study. The correlation analysis was done using Pearson's rank correlation method.

3.3.2. Software used

The above mentioned genetic indices were estimated in a PERL scripting language program developed by us to measure the CUB on the selected coding sequences of proto-oncogene, oncogene and tumour suppressor genes. All statistical analyses were carried out using the SPSS software. Cluster analysis (Heat map) of correlation coefficient of codons with GC_3 values of codons among the coding sequences were clustered using a hierarchical clustering method implemented in NetWalker software (Komurov *et al*. 2010).