Bioinformation
www.bioinformation.net
Volume 5
Issue 1
open access
Hypothesis

Analysis of SSR dynamics in chloroplast genomes of Brassicaceae family

Sumit G. Gandhi*, Praveen Awasthi, Yashbir S. Bedi

Plant Biotechnology Division and Systems Biology Division, Indian Institute of Integrative Medicine (Council of Scientific and Industrial Research), Canal Road, Jammu – 180 001, India; Sumit G. Gandhi - E-mail: sumit@iiim.res.in; *corresponding author

Received March 04, 2010 ; accepted April 09, 2010, published June 16, 2010

Abstract:
Simple sequence repeats (SSRs) are present abundantly in most eukaryotic genomes. They affect several cellular processes like chromatin organization, regulation of gene activity, DNA repair, DNA recombination, etc. Though considerable data exists on using nuclear SSRs to infer phylogenetic relationships, the potential of chloroplast microsatellites (cpSSR), in this regard, remains largely unexplored. In the present study we probe various nucleotide repeat motifs (NRMs) / types of SSRs present in chloroplast genomes (cpDNA) of 12 species belonging to Brassicaceae family. NRMs show a non-random distribution in coding and non-coding compartments of cpDNA. As expected, trinucleotide repeats are more common in coding regions while other repeat motifs are prominent in non-coding DNA. Total numbers of SSRs in coding region show little variation between species while considerable variation is exhibited by SSRs in non-coding regions. Finally, we have designed universal primers that yield polymorphic amplicons from all 12 species. Our analysis also suggests that amplicon length polymorphism shows no significant relationship with sequence based phylogeny of SSRs in cpDNA of Brassicaceae family.

Keywords: SSR, microsatellites, phylogenetic relationship, chloroplast DNA, *Brassicales*, coding DNA, non-coding DNA

Back ground:
SSRs or microsatellites are tandem repeats of mono-, di-, tri-, tetra-, etc. nucleotide motifs. They infest the genomes of most eukaryotes and often exhibit length polymorphism. The reversible length altering mutations are resultant of unequal crossing over and replication slippage. Relative conservation of the flanking regions, allow the variable length microsatellites to be used as locus specific, co-dominant, genetic markers across taxa. SSRs present in non-coding DNA were earlier thought to be non-functional entities, but now we know that they play distinct roles in genome organization, regulation of transcription, DNA recombination and repair, etc. **[1]** When present in protein coding DNA segments, their expansion or contraction can have huge impact on protein's function. Several human diseases have been linked to expansion of trinucleotide microsatellites within protein coding genes and have been dubbed trinucleotide repeat disorders. **[2]** Through "guilt by association", several microsatellite loci in plants have been linked to stress tolerance, disease resistance, domestication events and various agronomic traits. **[3]** Cp DNA has lesser percentage of non-coding component as compared to nuclear DNA, still SSRs are abundant in chloroplast genomes. In contrast to nuclear DNA markers that are inherited both from seed and pollen the cpDNA is inherited only through maternal route in angiosperms. It is considered a highly polymorphic marker that can be used to trace divergence through geographical isolation. **[4]** The present work finds the relative percentages of different SSR motifs and their distribution in coding and non-coding compartments of cpDNA in members of Brassicaceae family. Brassicaceae includes almost 338 genus comprising of about 3700 species exhibiting cosmopolitan distribution. **[5]** The family is also known as mustard family and members are mostly annual or perennial herbs. *Arabidopsis thaliana*, a commonly used 'model plant' also belongs to this family. Chloroplast genome sequences of 12 species of Brassicaceae are present in GenBank **[6]** and we have included all of them in our analysis of microsatellite loci. The study also tests the appropriateness of using data from cpDNA SSR length polymorphism as a genetic marker to plot phylogenetic relationships.

Methodology:
Mining of chloroplast genome sequences
cpDNA sequences of 12 species of Brassicaceae family, namely, *Aethionema cordifolium* (NC_009265), *Aethionema grandiflorum* (NC_009266), *Arabidopsis thaliana* (NC_000932), *Arabis hirsuta* (NC_009268), *Barbarea verna* (NC_009269), *Capsella bursa-pastoris* (NC_009270), *Crucihimalaya wallichii* (NC_009271), *Draba nemorosa* (NC_009272), *Lepidium virginicum* (NC_009273), *Lobularia maritima* (NC_009274), *Nasturtium officinale* (NC_009275) and *Olimarabidopsis pumila* (NC_009267) were downloaded from GenBank **[6]** (http://www.ncbi.nlm.nih.gov). DNA sequences for coding and non-coding DNA segments of cpDNA were downloaded from the Chloroplast Genome Database **[7]** (http://chloroplast.cbio.psu.edu).

Mining SSRs from cpDNA and Primer design
A standalone perl script was used to find SSR motifs. For monomers, the minimum repeat size was 10 nt, for dimers, minimum repeat size was 5 nt, for trimer to decamer, minimum repeat size was 3 nt. Both perfect SSRs and compound SSRs were detected. The maximum interruption size between compound SSRs was kept 5 nucleotides. SSRs were searched in full chloroplast genome as well as separate coding and non-coding regions for each species. About 200-400 nt sequences flanking the SSR was used in online tool primer3 **[8]** for designing primers. Parameters used for primer3 were: optimum primer size - 20nt, optimum annealing temperature - 59°C, optimum GC content – 50%.

SSR dynamics in coding, non-coding and complete cpDNA
The data generated from SSR mining was analyzed using Microsoft Excel®. Percentages of different types of SSR motifs were calculated and their occurrence in coding vs. non-coding DNA was determined. The number of SSRs in coding and non-coding DNA across species was represented graphically to appreciate the dynamics.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 16-20 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1
open access
Hypothesis

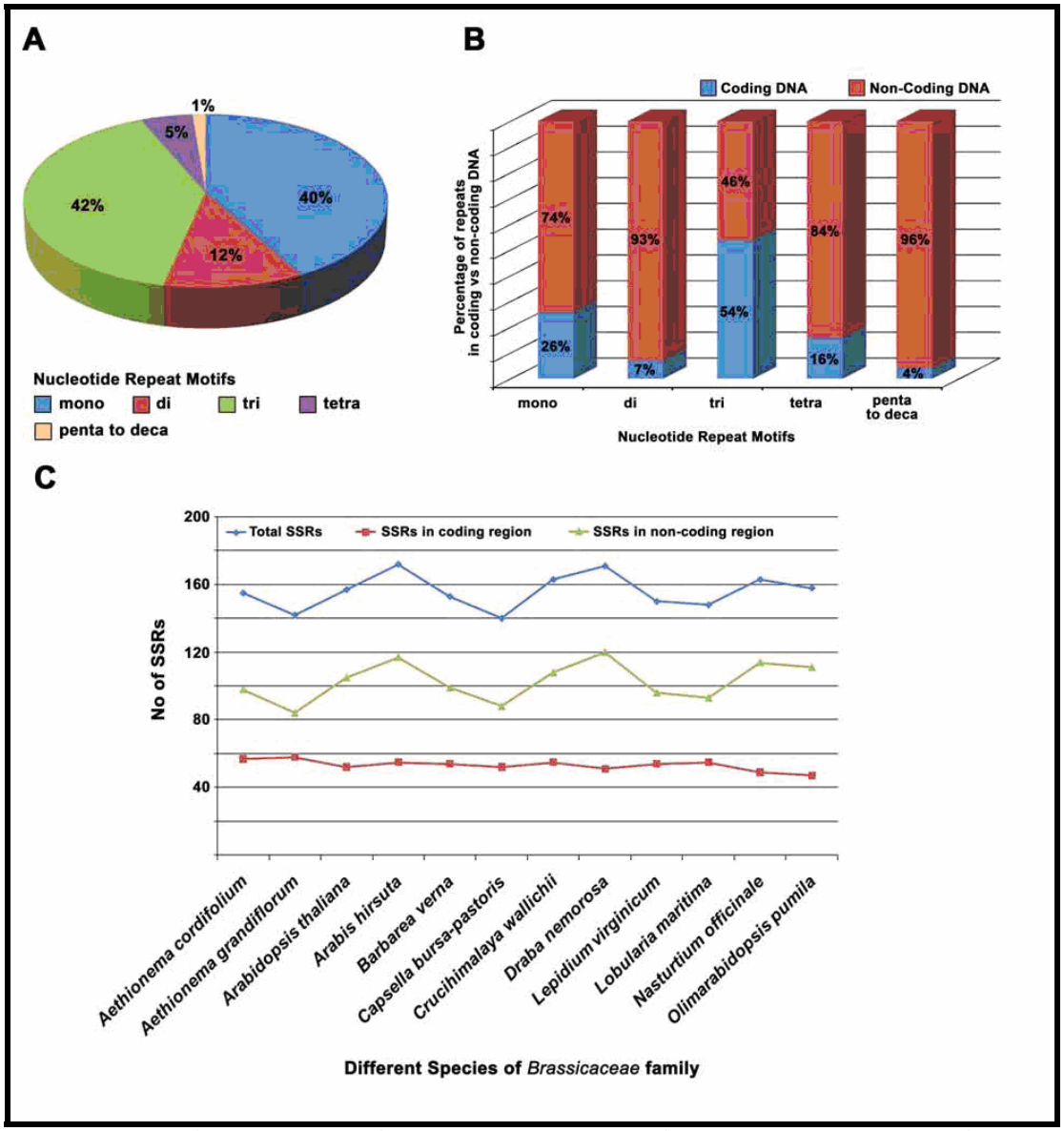

Figure 1: Distribution of nucleotide repeat motifs (NRM) in choloroplast genomes of 12 species belonging to Brassicaceas family,(a) pie chart revealing the percentage of different types of NRMs (SSRS) in Brassicaceae (b) Bar graph indicating percentage of various NRMS in coding or non-coding regions of chloroplast DNA (c) species wise distribution of total SSRS in coding or non-coding regions of chloroplast DNA note :-percentages have been rounded to nearest integer

Designing Universal primers & Virtual PCR on cpDNA :
200-400 nt flanking the SSRs were aligned and regions showing 100 % identity were used to design universal primers. These primers were tested using FastPCR **[9]** for amplification of SSR loci across cpDNA of all 12 species. Two sets of universal primers showing length polymorphism across all species were selected for further analysis.

Electrophoresis prediction and phylogenetic tree construction :
The amplicons generated by FastPCR from cpDNA of all species, using the selected universal primers, were loaded into CLC DNA workbench **[10]**.An *in-silico* gel electrophoretogram was plotted using the CLC DNA workbench. Length polymorphism data was used to generate a distance matrix, using the 'simint' module of NTSYSpc ver 2.2 **[11]**, and further clustered using UPGMA (Unweighted Pair Group Method with Arithmetic mean) algorithm for plotting a dendrogram. Amplicon sequence data was used to plot dendrogram on structural polymorphism basis. The amplicon sequences were aligned using ClustalW **[12]**. Neighbor-Joining clustering algorithm was used to plot a dendrogram. The robustness of the tree was tested using bootstrap analysis (1000 iterations).

Discussion:
Chloroplast genomes of 12 species of Brassicaceae family were analyzed for their total microsatellite content, distribution across coding and non-coding domains and relative percentages in different species. The data reveals interesting patterns of SSR distributions. Finally, using the designed universal primers, an attempt was made to use SSR length polymorphism for exploring phylogenetic relationships.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 16-20 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1 **Hypothesis**
open access

Types of SSR motifs in cpDNA of Brassicaceae :
Total numbers of different types of SSR motifs present in all cpDNA samples were estimated. Mononucleotide and trinucleotide motifs are most common and present in almost the same proportion (42% and 40% respectively). Motifs larger then penta nucleotide represent only 1% of total SSRs **(Figure 1a).** Twelve percent of the total NRMs are dinucleotide repeats, where AT/TA type repeats were predominant, in contrast to CG/GC type repeat. One of the reasons for this disparity may be that CG repeats are known to increase the stacking energy of DNA and can form Z-DNA in negatively super coiled regions, during transcription **[13]**.

Further, between the two most prominent repeat motifs, mononucleotide motif was overrepresented in non-coding DNA in contrast to the trinucleotide motif that was more common in coding DNA. Other NRMs were also present in larger numbers in non-coding DNA **(Figure 1b)**. Expressed sequence tags (ESTs) represent the coding portion of the genome.**[14]** Mining of publicly available EST sequences for SSRs have been performed by several groups and almost invariably, trinucleotide motif has been reported as most common **[15, 16]**, which corroborates with our results. This is expected as repeats other then trinucleotide motifs, would result in fram shift or null mutations, resulting in non-functional proteins, and hence would generally not be selected through evolution.

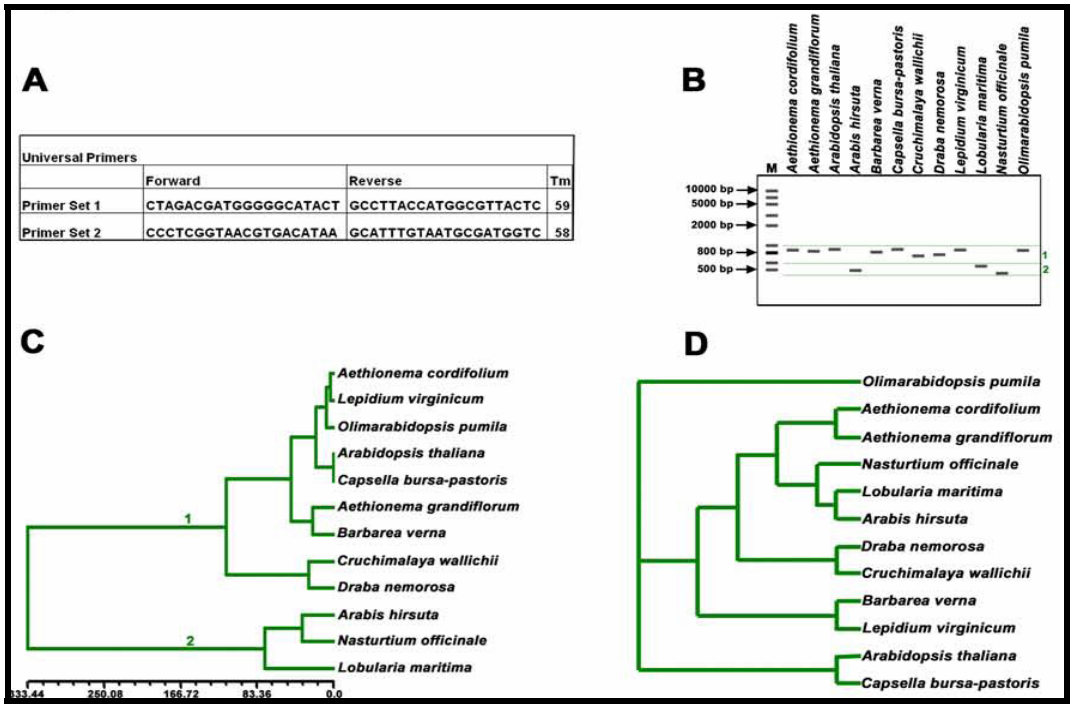

Figure 2: Microsatellite polymorphism in chloroplast genomes of 12 Brassicac eae species.(a) Universal primer sets exhibition amplicon length polymorphism (b) simulated DNA electrophoretogram of predicated amplicons using universal primer set 1 (b) simulated DNA electrophoretogram of predicted amplicons using universal primer set 1(c) phylogenetic tree based on amplicon length polymorphism using universal primer set1 (d) phylogenetic tree based on amplicon sequence polymorphism using universal primer set 1

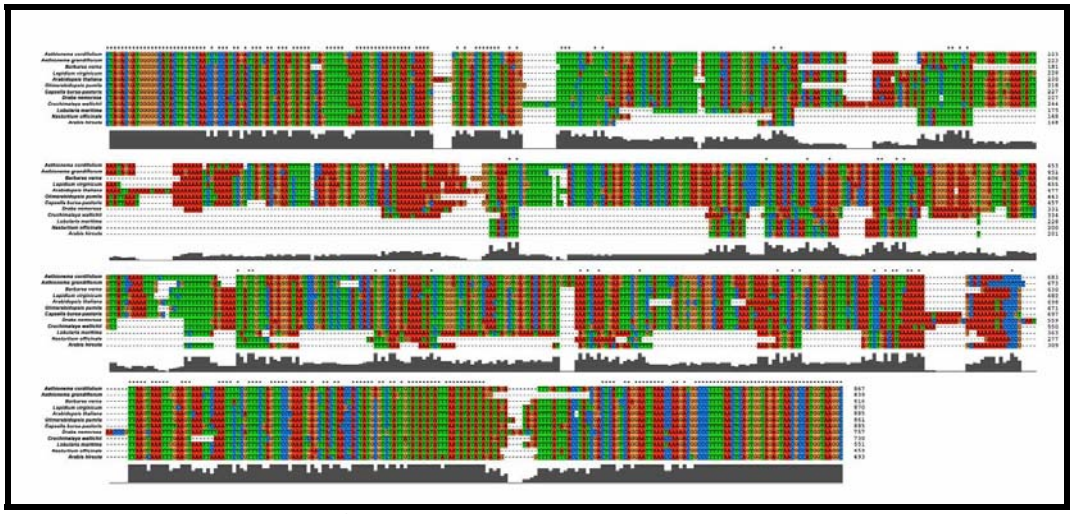

Figure 3 : Multiple sequence alignment of amplicon sequences using CLUSTAL W

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 16-20 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1 **Hypothesis**
open access

Distribution of SSRs in cpDNA of different species
Distribution of total numbers of SSRs present across the 12 species was tested. Total numbers of SSRs in complete cpDNA molecules show considerable variation across different species. Further analysis showed that this huge variation is mainly due to the differences in SSR content in non-coding regions. Coding regions of different cpDNA, in contrast, showed lesser variation in SSR numbers, across the species **(Figure 1c** & **Table 1** in Supplementary data). As a corollary from this analysis, it also appears that despite 51% of cpDNA (of selected species) encodes proteins, total number of SSRs in coding regions range only from 47 to 58, in different species, while total number of SSRs in non-coding region (about 49% of chloroplast genome) range from 84 to 120. Thus, on an average, non-coding compartment of cpDNA in Brassicaceae contains twice the number of SSRs present in coding DNA. Simillar results have been reported for mitochondrial and chloroplast genomes of rice **[17]**. SSRs are known to undergo cyclical expansion and contraction. These mutations are caused by recombination or by replication slippage **[1]**. Such changes would be more easily tolerated in non-coding DNA and this could be one of the reasons why more numbers of SSRs are observed in non-coding regions.

Microsatellite polymorphism across Brassicaceae:
Length polymorphism of SSRs can be easily assessed using simple polymerase chain reaction (PCR) assays, with primers designed on conserved flanking regions **[18]**. Two sets of universal primers, which yield variable length amplicons across the 12 Brassicaceae species, have been designed **(Figure 2a).** An electrophoretogram was simulated to demonstrate the extent of length polymorphism **(Figure 2b).** On basis of amplicon size distribution, two major classes emerged. These have been marked in the electrophoretogram, and are also represented in the dendrogram plotted on length variation basis **(Figure 2c).** In order to test whether sequence based phylogeny (structural polymorphism) correlates with the dendrogram plotted using length variation of SSR, we aligned the amplicon sequences in clustalW **(Figure 3**). Subsequently, clustering was performed and sequence based phylogenetic tree was constructed **(Figure 2d).** Our analysis suggests that SSR length polymorphism and structural polymorphism may not correlate for cpDNA of Brassicaceae. Similar results were reported for SSR data from mitochondrial genome of domestic animals **[19]** as well as for chloroplast genome in *Cucumis* species **[20]**.

Conclusion:
This work reveals the distribution of different types of SSR in coding and non-coding compartments of 12 different species of Brassicaceae family. Out results indicate that, in general, SSRs have more preponderance in non-coding segments of cpDNA in contrast to the coding regions. However, triplet repeats are more prevalent in coding regions, as expected. Considerable variation in numbers of SSR is observed in non-coding regions of cpDNA across different species. Attempt to use amplicon length polymorphism to construct phylogenetic relationships did not yield results that were in complete congruity with amplicon sequence based phylogeny.

Acknowledgement:
Financial assistance provided by institutional CSIR grant is deeply acknowledged. The authors are thankful to the Director, IIIM for providing facilities for this work.

References:
[1] YC Li *et al. Mol Ecol* **11**:2453 (2002) [PMID:12453231]
[2] RL Stallings, *Genomics* **21**:116 (1994) [PMID:8088779]
[3] C Ruan, *African Journal of Biotechnology* **9**:573 (2010)
[4] J Provan *et al. Trends Ecol Evol* **16**:142 (2001) [PMID:11179578]
[5] MA Koch & K Mummenhoff, *Plant Syst. Evol.* **259**:81 (2006)
[6] DA Benson *et al. Nucleic Acids Res* **34**:D16 (2006) [PMID:16381837]
[7] L Cui *et al. Nucleic Acids Res* **34**:D692 (2006) [PMID:16381961]
[8] Primer3.http://frodo.wi.mit.edu/primer3/
[9] Fast PCR http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm
[10] CLC DNA Workbench http://www.clcbio.com/index.php?id=27
[11] NTSYSpc. http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html
[12] JD Thompson *et al. Nucleic Acids Res* **22**:4673 (1994) [PMID:7984417]
[13] S Rothenburg *et al. Proc Natl Acad Sci U S A* **98**:8985 (2001) [PMID:11447254]
[14] MA Marra *et al. Trends Genet* **14**:4 (1998) [PMID:9448457]
[15] T Asp *et al. BMC Plant Biol* **7**:36 (2007) [PMID:17626623]
[16] L Zhang *et al. Bioinformatics* **20**:1081 (2004) [PMID:14764542]
[17] P Rajendrakumar *et al. Bioinformatics* **23**:1 (2007) [PMID:17077096]
[18] J Koreth *et al. J Pathol* **178**:239 (1996) [PMID:8778326]
[19] SK Shakyawar *et al. Bioinformation* **4**:158 (2009)
[20] SM Chung *et al. Genome* **49**:219 (2006) [PMID:16604104]

Edited by P. Kangueane
Citation: Gandhi *et al.* Bioinformation 5(1): 16-20 (2010)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 16-20 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1 *open access* **Hypothesis**

Supplementary material:

Table 1 :

SSR	*Aethionema cordifolium*		*Aethionema grandiflorum*		*Arabidopsis thaliana*		*Arabis hirsuta*		*Barbarea verna*		*Capsella bursa-pastoris*	
Type	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding
Mono-	19	43	19	41	17	52	16	54	20	50	16	40
Di-	1	15	2	12	1	15	2	22	1	16	2	13
Tri-	35	32	36	26	33	27	35	30	32	29	33	28
Tetra-	1	8	1	5	1	8	2	10	1	4	1	5
Penta-	0	0	0	0	0	2	0	1	0	0	0	2
Hexa-	1	0	0	0	0	1	0	0	0	0	0	0
Hepta-	0	0	0	0	0	0	0	0	0	0	0	0
Octa-	0	0	0	0	0	0	0	0	0	0	0	0
Nova-	0	0	0	0	0	0	0	0	0	0	0	0
Deca-	0	0	0	0	0	0	0	0	0	0	0	0
Total	57	98	58	84	52	105	55	117	54	99	52	88

SSR	*Crucihimalaya wallichii*		*Draba nemorosa*		*Lepidium virginicum*		*Lobularia maritima*		*Nasturtium officinale*		*Olimarabidopsis pumila*	
Type	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding	Coding	Non-Coding
Mono-	19	52	17	57	14	46	19	43	13	58	14	48
Di-	1	18	0	24	1	14	2	14	1	16	1	18
Tri-	34	26	33	28	38	29	33	27	32	33	31	34
Tetra-	1	7	1	7	1	7	1	4	3	6	1	9
Penta-	0	2	0	2	0	0	0	1	0	1	0	2
Hexa-	0	0	0	0	0	0	0	4	0	0	0	0
Hepta-	0	1	0	2	0	0	0	0	0	0	0	0
Octa-	0	1	0	0	0	0	0	0	0	0	0	0
Nova-	0	0	0	0	0	0	0	0	0	0	0	0
Deca-	0	1	0	0	0	0	0	0	0	0	0	0
Total	55	108	51	120	54	96	55	93	49	114	47	111

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 16-20 (2010)