BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(16)

Detecting motifs and patterns at mobile genetic element insertion site

Bhuvan Bhaskar Dev[1, 2], Aman Malik[1] & Kamal Rawal[1]*

[1]Department of Biotechnology, Jaypee Institute of Information Technology, Noida 201307, UP, India; [2]Infosys Ltd, Bangalore, Karnataka, India; Kamal Rawal – Email: kamal.rawal@jiit.ac.in; *Corresponding author

Received July 17, 2012; Accepted July 23, 2012; Published August 24, 2012

Abstract:
Mobile genetic elements (MGEs) occupy major proportion of eukaryotic genomes and are present in significant numbers in prokaryote genomes also. Here we report a new method which extracts a motif at the site of insertion of MGE using tools such as DNA SCANNER. The flanking region of the insertion site is extracted and is analyzed in DNA Scanner for physiochemical properties like protein-interaction measures, energy profiles as well as structural parameters. In case significant signals are observed, the most frequently occurring n-mer (5<n<20) is identified. We observed signals between 9-12 base pairs (bps) upstream in pre-insertion site of Alu element in Human and most frequently occurring motif is found to be TTAAAA. The similar signals and motif is observed at insertion site of B1 element. In lower eukaryotes such as *E. histolytica,* signals for EhSine1 are found at around 5 bps upstream of insertion and most frequently occurring motif is found to be AAGGT and TCGAA. Signals for Ty3 element in *S. cerevisiae* are found at 0-3 bps upstream of tRNA, and most frequent motif is GTTCGA (6 bps), GGTTCGA (7 bps) and GGTTCGAT (8 bps). P-element of Drosophila showed remarkable dyad peaks suggesting palindromic site of insertion.

Background:
Presence of mobile genetic elements is a widespread phenomenon in various taxonomical groups. For every genus there are families of MGE, which are classified based on sequence similarity, origin and mode of transposition. For example in Human genome Alu is the most successful MGE in terms of copy number whereas LINE-1 occupies most of the human genome. In mouse- B1 and L1, in *E. coli* – Tn7, in Drosophila – P element, in *E. histolytica*- EhLINE, EhSINE etc. are some of the widely studied elements. The role of mobile genetic elements has always been debatable in relation to their mode and site of insertion and their effect on biology of organism. Some of the functions which have been documented are in the context of gene inactivation, transduction of genomic sequences, regulation of gene expression and genomic expansion.

Identification of specific motifs or patterns at insertion site of MGE is one of the interesting problems studied by several researchers in the past. Patterns or motifs are found in number of important contexts in biological systems which include DNA sequences, protein sequences, RNA etc. Motifs such as TATA box, β-α- β, Greek key, HTH etc are some of the well studied examples in the literature. The presence of specific patterns or motifs like TTAAAA at Alu insertion sites was reported by Jurka **[1]**. This motif was validated by the several numbers of experiments which find that Alu insertion occurs with highest frequency in site downstream of TTAAAA motif **[2]**. Presence of such patterns has also been discovered in P-element of *D. melanogaster* **[3]** and EhSINE1 of *E. histolytica* **[4]**.

Here, we present a method of analyzing the DNA sequence of insertion site based on its secondary features. These features can be broadly divided into structural parameters like Bendability and Propeller twist; energy profiles like Stacking energy, duplex stability free-energy and stabilizing energy and protein interaction parameters like Protein induced deformability, Nucleosomal positioning and Bending Stiffness; sequence based parameters like A-philicity, AT-rule, B to A trimeric form of DNA, C-rule, G-rule and T-rule. We hypothesize that transposition machinery first recognizes donor DNA based on these structural parameters and then the specific nucleotides such as TTAAAA. We applied our approach on wide variety of examples which includes pre-insertion loci of Alu element of Human, B1 (type of SINE) element of mouse, P-element in *Drosophila melanogaster*, EhSINE1 of *E. histolytica*, Ty3 in *S. cerevisiae* and Tn7 in *E. coli and D. desulfuricans.*

Previously our group has identified several sets of signals at insertion sites of various MGEs **[5]**; for example in Alu element at around 9-12 base pairs upstream of insertion site. During this study we found that TTAAAA motif is in highest frequency at flanking region of site of insertion of Alu. Set of experimental studies have independently verified these results **[1, 2]**. Encouraged by the results, we applied this approach to other examples from various taxonomic groups and also attempted to identify most frequently occurring motif in the flank of MGE insertion sites showing distinct physiochemical signals.

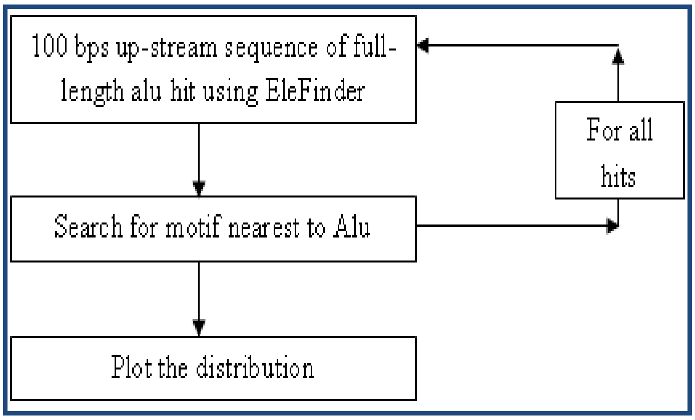
100 bps up-stream sequence of full-length alu hit using EleFinder
For all hits
Search for motif nearest to Alu
Plot the distribution

Figure 1: Flowchart to plot the distribution of motif in upstream sequences of the Alu.

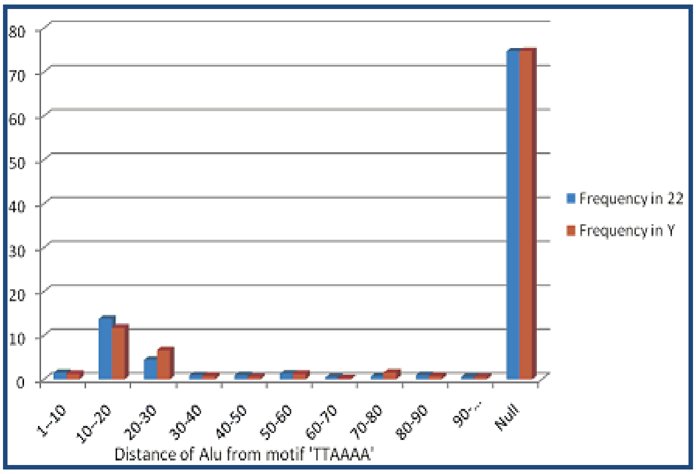

Figure 2: Distribution plot for TTAAAA motif, upstream of full length Alu.

Methodology:

Identification of the most frequently occurring motif in flank of Alu insertion site

To find motifs within 10-20 base pairs upstream of Alu, the full length hits are selected using the tool EleFinder **[5]**. The flanks (100 base pairs) are retrieved using Perl scripts and position of occurrence of motif nearest to insertion site is generated **(Figure 1).** Ten highest frequency motifs from the previous work **[1]** are analyzed in detail **(Figure 2).**

Retrieving Genomes and Repeat sequences

The sequence of human chromosome 22, X and Y, mouse chromosome 16, X and Y and *E. histolytica* genome are retrieved from ftp server of NCBI. *E. coli* and *D. sulfuricans* genome are retrieved from NCBI, accession no. NC_011601.1 and NC_007519.1 respectively. For Ty3 insertions in *Saccharomyces cerevisiae* genome, we selected tRNAs and flanking region. These are downloaded from genomic tRNA database **[6]**. Alu and B1 sequences are downloaded from RepBase **[7]**. EhSINE1 sequence was taken from our previous work **[8]**.

Finding insertion site

There are two types of insertion sites that can be used- one which are *in-silico* and the other *in-vitro*. While *in-vitro* data gives the actual insertion sites, *in-silico* methods are designed to give potential insertion sites and corrected statistically to give results close to as would be obtained *in-vitro.* We used EleFinder **[5]** to find the insertion site of Alu in human, B1 in mouse and EhSine1 in *E. histolytica.*

For elements Tn7 and Ty3, insertion is site-specific. Tn7 inserts downstream of glmS at 3` end **[9, 16]** and Ty3 inserts upstream of tRNA gene **[10]**. These are the most probable and highly specific site, thus are located directly. Tn7 insertion site in *E. coli* is extracted from EcoGene database **[11]**. It consists of 100bps of end of glmS and 100 bp downstream of it. Since glmS is conserved among bacteria, insertion site in *D. sulfuricans* is taken by identifying the glmS (of *E. coli)* in the bacteria by Blast and extracting 200 bps in similar way as *E. coli.* Ty3 insertion sites are taken from Genomic tRNA database, tRNA is removed and region upstream of it, 1500 bps and downstream of it, 1500 bps is used for analysis. Results of Elefinder for Alu and B1 are summarized in the **Table 1 (see supplementary material).** P-element insertion sites consisted positive strand dataset of 51 bp as described in Linheiro *et al* **[12]**.

Computational analysis of pre insertion sequences

For each of the features discussed above, a graphical profile is generated in sliding window along a sequence. DNA scanner **[5]** for Alu upstream sequences is run with window size of 5 and start position of 0, representing insertion site. The parameters are same for Chromosome 22, X and Y and B1 insertion sites in mouse chromosome 16, X and Y. EhSINE1 analysis is done using window size of 5 and start position as 0 representing insertion site. For *E. coli* and *D. sulfuricans* start position is taken as -100, 0 represents nucleotide at the end of glmS and window sizes were selected from 15 to 25; 20 is considered to be optimal size. For Ty3 window size is taken as 200 and start position as -1500, 0th position representing end of 5` upstream region of tRNA. For P-element window size is 5 and start position as -25, 0 marks the insertion site.

Finding motifs in the region around signal

After signals are found, the nucleotide sequence around extrema was used for finding the highest occurring motif for a given MGE. This was done using RSAT suite of tools **[13]**, under which we used 'Oligo Analysis'. We used following options–'mask non-DNA', 'prevent overlapping matches', 'count on single strand only', and results were displayed as 'Pattern Count Distribution'. The result was sorted based on occurrence of one instance of motif in a sequence. This is because the presence of just one motif per sequence is required for identifying the target by transposition machinery. Even if similar motif is present more than once in a sequence it will not affect the transposition though it may affect the site of transposition by few base-pairs.

Results:

(A) Distribution of motifs in 10-20 bps upstream of Alu insertion site

Jurka reported presence of n-mers (n<5<20) or motifs such as TT`AAAA in upstream sequence of around 400 examples of Alu sequences **[1]**. The first task was to find out whether such motifs are also present in several thousand full length copies present in context of whole human genome. Our analysis revealed that out of the ten most frequently occurring motif according to previous work, TTAAAA has the highest frequency **Table 2 (see supplementary material)**, as reported earlier **[1]**. Our approach extended the work at genome wide level. TTAAAA is found in highest frequency at 10-20 bp upstream of insertion site **(Figure 2).** Apart from this, it is observed that CTAAAA is not present on Y chromosome insertion sites and TAAAAA motif is the second most frequent occurring motif instead of TTAAGA **Table 2 (see supplementary material).**

As a representative case we report 1464 examples in chromosome 22, 3314 in X chromosome and 391 in chromosome Y in the current work. Using this dataset of pre-insertion loci, full length alu insertions are characterized using sliding window method of DNA Scanner tool **[5]**. We identified 14 physiochemical properties in terms of energetics and structural parameters, described above and obtained extrema at the site of insertions. The method is also applied to large number of elements in other organisms which include *E. coli*, *S. cerevisiae*, *E. histolytica* and *Drosophila melanogaster*. At sites of extrema we found presence of certain 'words' or substrings of DNA sequences which are found in higher frequency than others. Incidentally in some cases, these substrings are found to be experimentally determined endonuclease nicking sequences or consensus sequence. These words are mostly present in flank of insertion sites correlating with physiochemical signals identified by the DNA Scanner.

(B) Signal in pre-insertion sites

(1) Signals and motif in 5` upstream region of Alu insertion sites in Human and B1 insertion sites in Mouse

Alu elements are non-autonomous retrotransposons that mobilize in a copy and paste fashion. It is proposed that the first nick at the site of insertion is often made by the L1 endonuclease at the TTAAAA consensus site **[14]**. B1 element of mouse is similar to Alu element of human, both of which originated from 7SL RNA, a signal recognition particle involved in translation initiation of eukaryotic secreted proteins. Like in human, B1 occupy a significant portion of mouse genome **[15]**. Unlike, Alu which is a dimer, B1 is a monomer- 140 bps in length but like Alu they are type of SINEs and require reverse transcriptase encoded by L1 for their transposition.

We show results of 1464 Alu and 1861 B1 examples as a representative case **Table 3 (see supplementary material).** We found that extrema lie in the region of 9-12 bp upstream for Alu and 15-18 bp upstream for B1. The minima are denoted as 'D' and maxima as 'U' in Table 3 along with other details such as position of signal and value of the parameter. The T-rule shows major deviation, its extrema being at -20 bp for Alu and at -50 for B1. The site of extrema for Alu coincided with the region where TTAAAA is found in highest frequency as compared to other motifs **(Table 2 & Figure 2)**. Using RSAT **[13]**, we found out that TTAAAA is the most frequent occurring motif for both the elements. The presence of signals in the same region as well as utilization of similar motifs indicates that B1 and Alu tend to retrotranspose using similar mechanism.

(2) Signals and motifs at 5` upstream regions of EhSINE1 insertion sites in E. histolytica

E. histolytica occupies a unique position in evolution between prokaryotes and eukaryotes. The EhLINEs and EhSINEs account for 6-8% of host genome **[8]**. EhSINE1 is the most frequently occurring MGE in *E. histolytica* genome. We used 64 pre-insertion loci to screen 50 bp flanks. The unique physiochemical signals are found in the region of around 5 bp upstream i.e. near insertion site **(Figure 4).** The region has relatively high bendability, low stacking energy and AT frequency. The results are similar to work done previously **[4]**. Most frequent 5-bp motif using RSAT **[13]** is found to be AAGGT.

(3) Signals and motif at insertion site of P-element

The P-element is 2907 bps long and contain transposase gene. Non-autonomous P-element exists due to some internal deletions which makes transposase gene inactive. Such element can move with the help of autonomous P-element. The mode of transposition is via cut-and-paste sequence of DNA transposon without involving any RNA intermediate. The donor element is excised and inserted into target site leading to target site duplication. We analyzed 5091 such insertion sites. Each site is extracted and centered around the insertion site with 25 bps on each end. The signals are very sharp showing perfect dyads centered on the insertion site **(Figure 5).** All the 14 rules are positive. This suggests the presence of palindromic motif as has been postulated previously **[12]**. Most frequent 8-bp motif is found to be CTCTCTCT.

***(4) Signals* at insertion site of Tn7**

Tn7 is 14kb element carrying 5 genes apart from 2 resistance genes - Trimethoprim (Tm^R) and Streptomycin/ Spectinomycin (Sm^R, Sp^R). The Tn7 transpositions have been studied extensively because of the high efficiency of transposition and also because of the specificity of its insertion site called attTn7. This site is located downstream of the coding region of glmS gene, in its transcriptional terminator.

Tn7 encodes 4 proteins- TnsA, TnsB, TnsC and TnsD which help in insertion. Site specific protein TnsD recruits to attTn7 which help in directing other Tns proteins and ends of transposon to the site. TnsA and TnsB constitute transposase which nicks and inserts the transposon **[9]**. The actual site of insertion is such that glmS remains unaffected. The gene glmS encodes glucosamine synthetase- required for cell wall synthesis and is conserved in bacteria and the same site, attTn7 has been found in many different bacteria like *D. sulfuricans* **[16, 19]**. Site specific insertion at attTn7 site requires binding of TnD to region -25 to -55 in glmS, the actual site of insertion is away from this site. In our data we find unique patterns at the end of glmS **(Figure 6 A, C & E).** For *D. sulfuricans* the signals lie just downstream of glmS gene i.e. 0th position suggesting that insertion in this host also does not affect the gene **(Figure 6- B, D & F).**

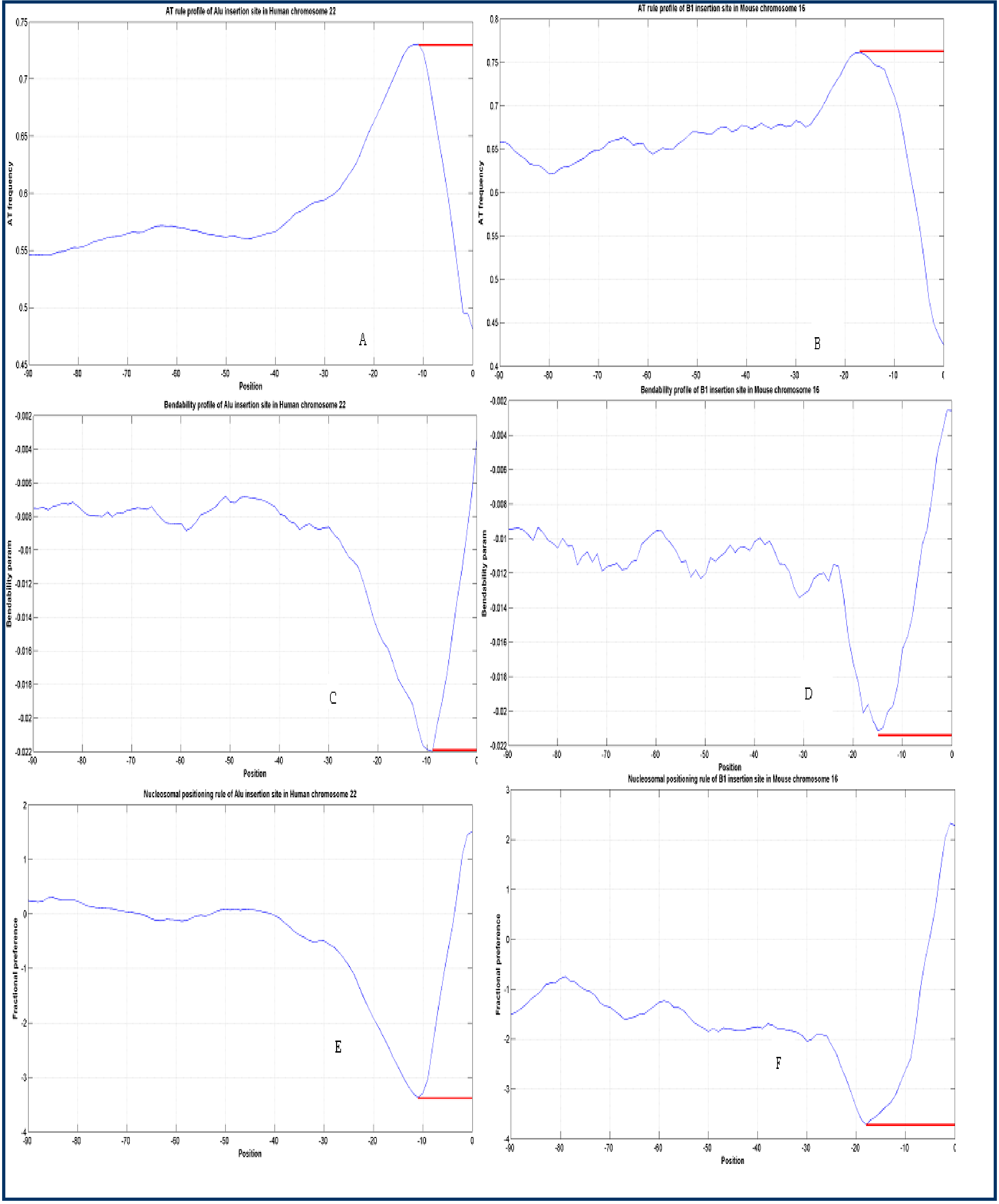

Figure 3: Various signals upstream of the insertion sites of alu in human chromosome 22 and mouse chromosome 16, for **(A, B)** AT rule, **(C, D)** Bendability, **(E, F)** Nucleosomal Positioning. The y axis represents value of the property and the x-axis gives the relative position with respect to the insertion site (taken to be 0). Red line indicates distance of insertion site from the signal.

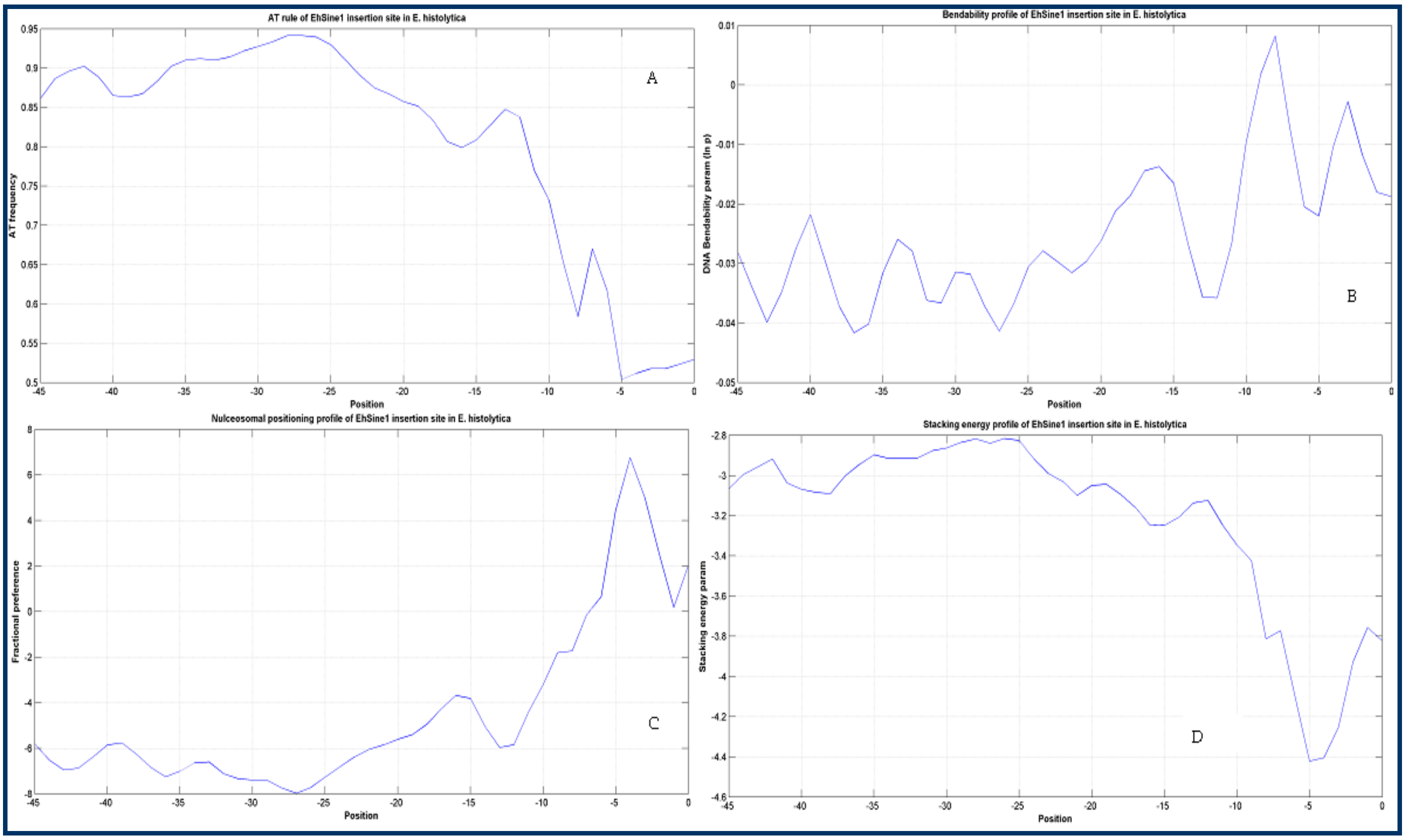

Figure 4: Various signals upstream of the insertion sites of EhSine1, for **(A)** AT rule, **(B)** Bendability, **(C)** Nucleosomal Positioning and **(D)** Stacking Energy. The y axis represents value of the property and the x-axis gives the relative position with respect to the insertion site, taken to be 0.

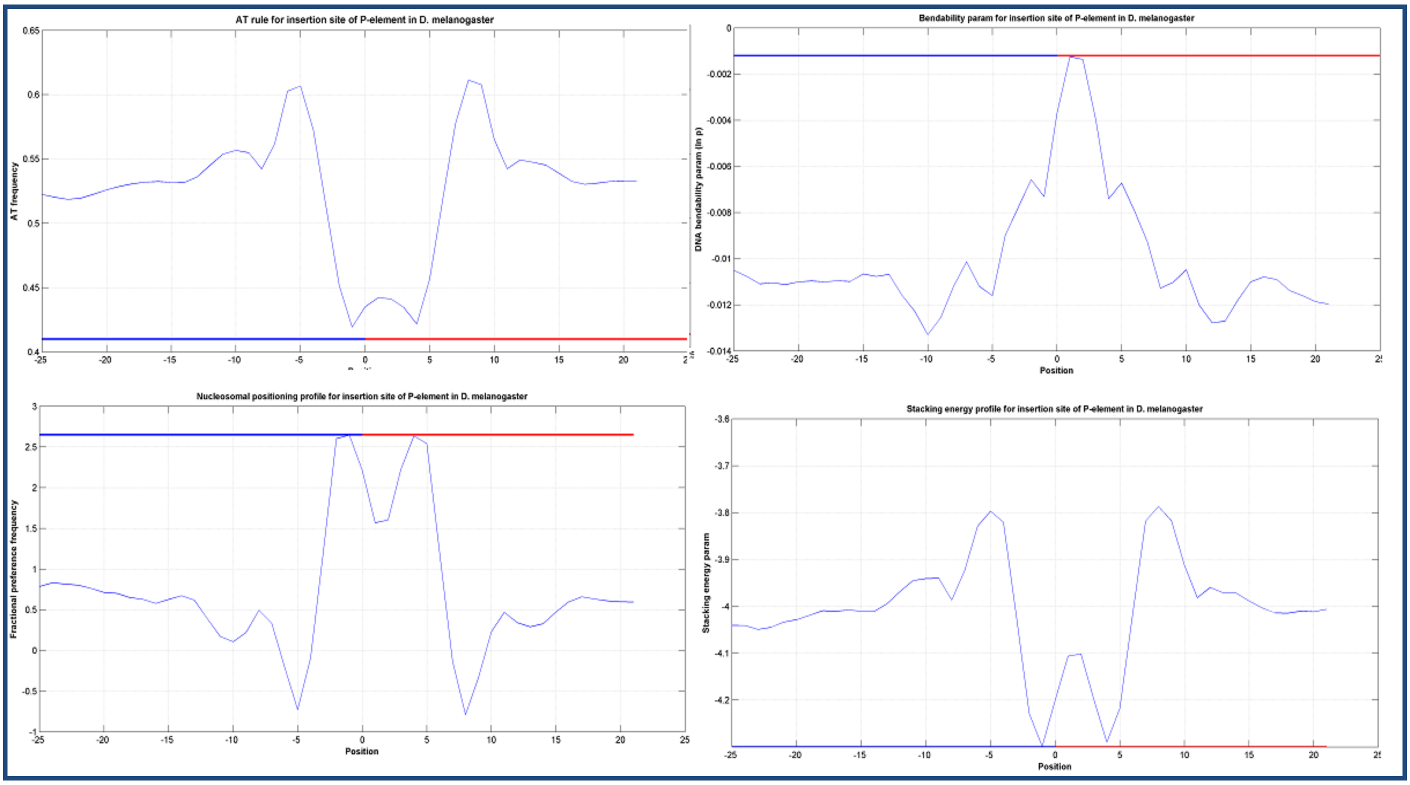

Figure 5: Various signals of the insertion sites of P-element, for **(A)** AT rule, **(B)** Bendability, **(C)** Nucleosomal positioning energy and **(D)** Stacking energy. The y axis represents value of the property. Red line indicates 5` upstream and blue indicates 3` downstream of insertion site, 0 being insertion site

(5) Signals at insertion site of Ty3 in S. cerevisiae

Ty3 is one of the five Ty element found in the host organism. They have insertion specificity and ability to transit the nuclear membrane that is a characteristic of retroviral elements. Their site of integration is the site of transcription initiation by RNA Polymerase III **[17]**. Around 3000 bps sequence in chromosome 4 and chromosome 7 are analyzed. Very sharp signals are observed at 5\` upstream of tRNA, extrema is observed at around 0th position **(Figure 7).** Results of chromosome 4 are discussed here as a representative case. Most frequent 6-mer motif in chromosome 4 and 7 is found to be GTTCGA and TTTTTT respectively. Other motifs of 7 bps and 8 bps are shown in **Table 4 (see supplementary material).** Results other than given here can be accessed at: https://www.dropbox.com/sh/ctufqucoq6dl24l/0mp6KMa-R_.

Discussion:

Our work compiles and identifies several motifs and patterns in wide variety of the examples drawn from *E. coli*, *E. histolytica*, Human genome etc. It identifies some of the previously reported motifs such as TTAAAA in Alu in human and the same motif for B1 element in Mouse genome. It also misses out or reports new motifs in case of *E. histolytica* genome. The unique and clear signals are seen in yeast genome w.r.t. tRNA sequences showing strong signals at their sites of insertion. The presence of Dyad motif at P element is also important aspect of our work.

Alu occupies 10 % of human genome and is the most actively transposing element in human genome. It is found that TTAAAA is the most frequently occurring motif upstream of Alu insertion sites. The distribution pattern showed that it occurs most frequently in 10-20 bps upstream **(Figure 2).** The results are encouraging as they are in agreement with previous work done **[1, 2, 18]**. It was shown in previous work that apart from TTAAAA, other motifs like TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, TCAAGA, AAAAAA, TTTAAA, TAAAAA, and GTAAGA etc. were also present 15-16 upstream of insertion loci **[1]**. Thus the distributions of these patterns are studied, in respect of their occurrence within 10 bp interval. Though identified across 100 bp upstream of insertion sites, they are majorly found 10-20 bp upstream regions except CTAAAA **Table 2 (see supplementary material).**

Target DNA chromatinization affects L1 endonuclease **[18]** activity therefore different parameters such as nucleosome based parameters as well as physiochemical properties are characterized at insertion sites. The signals observed for Alu pre-insertion loci in human chromosome 22, are present 9-12 bp upstream of insertion sites. The adenine excess at 10 bp upstream of insertion site relative to other regions is a marked characteristic of Alu pre-insertion loci. A region of low bendability is observed in the pre-insertion locus which gives a sharp dip at 9 bp upstream. Low denaturation energy and higher free energy at -9 to -12 signifies that endonuclease identifies such region for retrotransposition. Also high stacking and stabilizing energy signals show that such regions are preferred for insertion. Apart from chromosome 22, chromosome X (3314 insertion sites) and chromosome Y (391 insertion sites) are also analyzed. We found similar signals as well as motifs.

B1 is a monomer and seem to depend on L1 for its movement. This indicates that B1 has similar preference of insertion sites as seen in Alu. B1 shares structural similarity with Free Left Arm Monomer (FLAM) of Alu**.** We found that the most of the signals are observed 15-18 bp upstream of insertion sites in 1861 insertion examples of mouse as compared to human where signals were mostly seen 9-12 bp upstream **(Figure 3)**. The pre-insertion loci of B1 are characterized by relatively high stacking and stabilizing energy and also adenine excess. Most frequently occurring motif is TTAAAA in all chromosomes of mouse analyzed- 16, X and Y. This seems to validate the hypothesis that both Alu and B1 share similar mode of retrotransposition.

The EhSINE1 of *E. histolytica* is also an example of SINE, whose mode of retrotransposition is similar to Alu. It requires EhLINE1 for its transposition machinery. Among all the SINEs found in host organism, its frequency of insertion is the highest and hence it is used as a representative element. The signals are found in region up to -5 bp of insertion sites. The positive signals found in our work and previously reported **[4]** are same. The pre-insertion loci have adenine and cytosine excess, high bendability, low denaturation energy and duplex stability free energy **(Figure 4).** We analyzed 50 bp upstream of insertion site in RSAT **[13]** for the most frequently occurring motif and found - AAGGT, TCGAA, GAAGG, AGATC, and ATCGA occurring in highest frequency. This is different from previous reported motif- GCATT. This can be attributed to different amount of data-set analyzed by us as compared to in-vitro studies done so far.

P-element is a classic example of DNA transposon which moves by cut and paste mechanism. The extracted insertion site consists of 25 bp of 5\` and 25 bp of 3\` end of insertion site, 0 representing insertion site. The signals are perfect dyads **(Figure 5).** The pre and post insertion sequences show adenine excess whereas insertion site gives a sharp dip. High bendability value i.e. low rigidity and low bending stiffness at insertion site reflect their role in transposition machinery. We tried to identify 8 bp motifs and found it to be CTCTCTCT. The results are in complete agreement with the dyads peak which shows that insertion site is palindromic. Other frequently occurring motif found are AGAGAGAG and TCTCTCTC which are also palindromic. Also the signals shows region of adenine, cytosine, thymidine and guanine excess consistent with observed motif composition **(Figure 5).**

Tn7 is one of the site-specific transposon which inserts downstream of transcriptional terminator of glmS gene by identifying a region -25 to -55 bp at its 3\` end. This suggests that region is highly specific for various parameters that we are studying using DNA Scanner. Downstream region of glmS has cytosine and guanine excess and low adenine content and has high bendability. The insertion site has 13 out of 14 parameters positive, which means that certain characteristic features decide the insertion sites of Tn7 **(Figure 6)**. The same set of signals is obtained for *Desulfovibrio desulfuricans* G20 which previously has been shown to have attTn7 site **[16]**. Since we analyzed only *E. coli* and *D. sulfuricans* genome for this work, motif identification is not done. To find the motif, we need to analyze more bacteria like *Shigella sonnei Ss046, Raoultella terrigena, Enterobacter cloacae, Pseudomonas aerugenosa, Shewanella putrefaciens* CN-32 etc **[19]** that show Tn7 transposition.

Ty3 of *S. cerevisiae* is one of the three actively retrotransposing element. Ty3 shows site-directed insertion, with reported insertion 0-2 bp upstream of tRNA **[17]**. All the 14 parameters are positive and showed signals upstream of 5` end of tRNA. The pre-insertion region shows low adenine and high guanine, cytosine and thymidine content. It has low Bendability and thus high bending stiffness **(Figure 7)**. Motif analysis of 200 bps sequence around the insertion site is done. Since no previous study is done in this regard we extracted 6, 7 and 8 bp motif. The most frequently occurring motifs have GTTCGA as common pattern which is consistent with high T, C, and G content. Also the next most frequently occurring motif is thymidine rich **Table 4 (see supplementary material).**

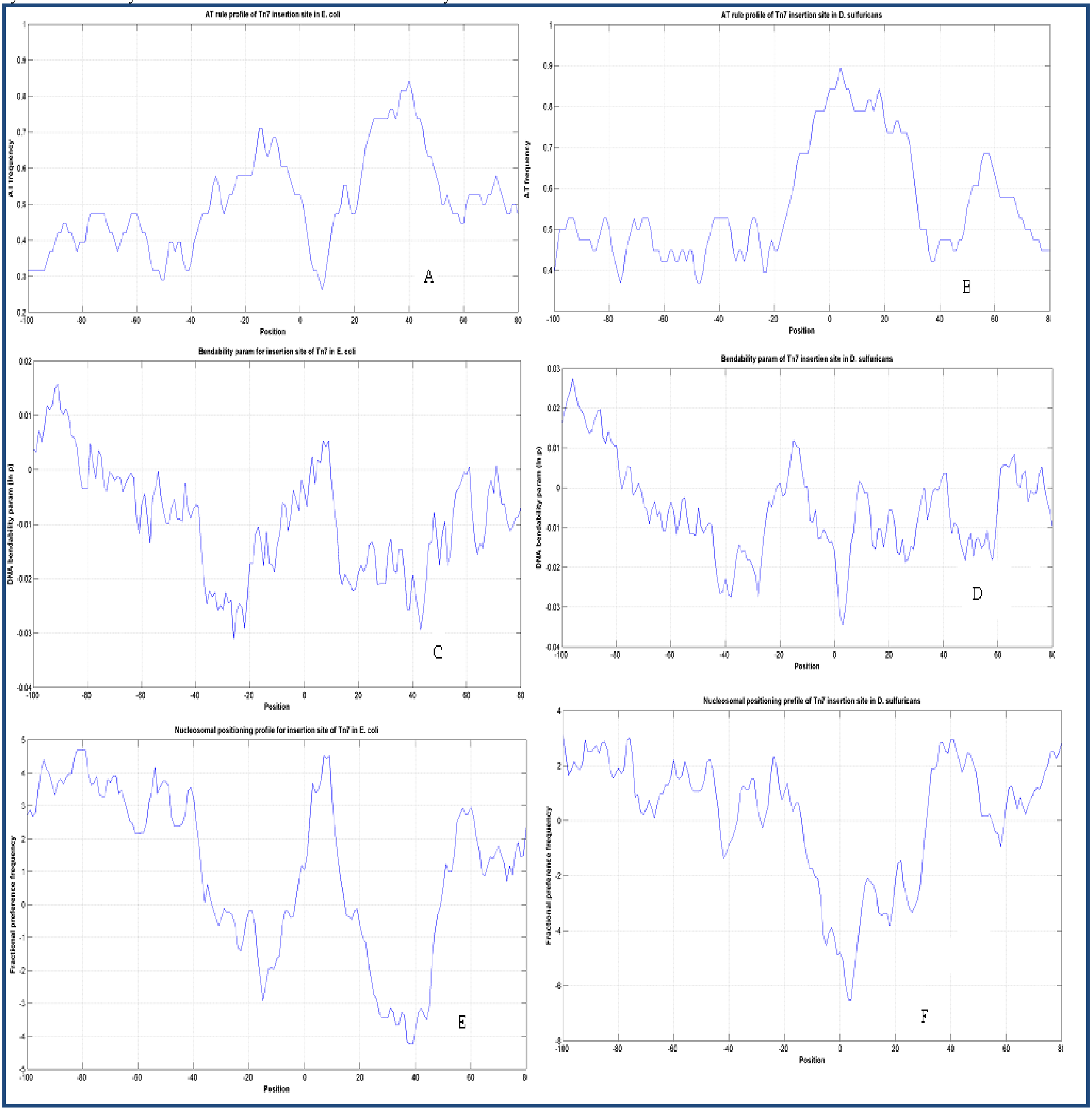

Figure 6: Various signals at insertion sites of Tn7 in *E. coli* and *D. sulfuricans*, for **(A, B)** AT rule, **(C, D)** Bendability, **(E, F)** Nucleosomal Positioning. The y axis represents value of the property and the x-axis gives the relative position with respect to the insertion site, taken to be 0.

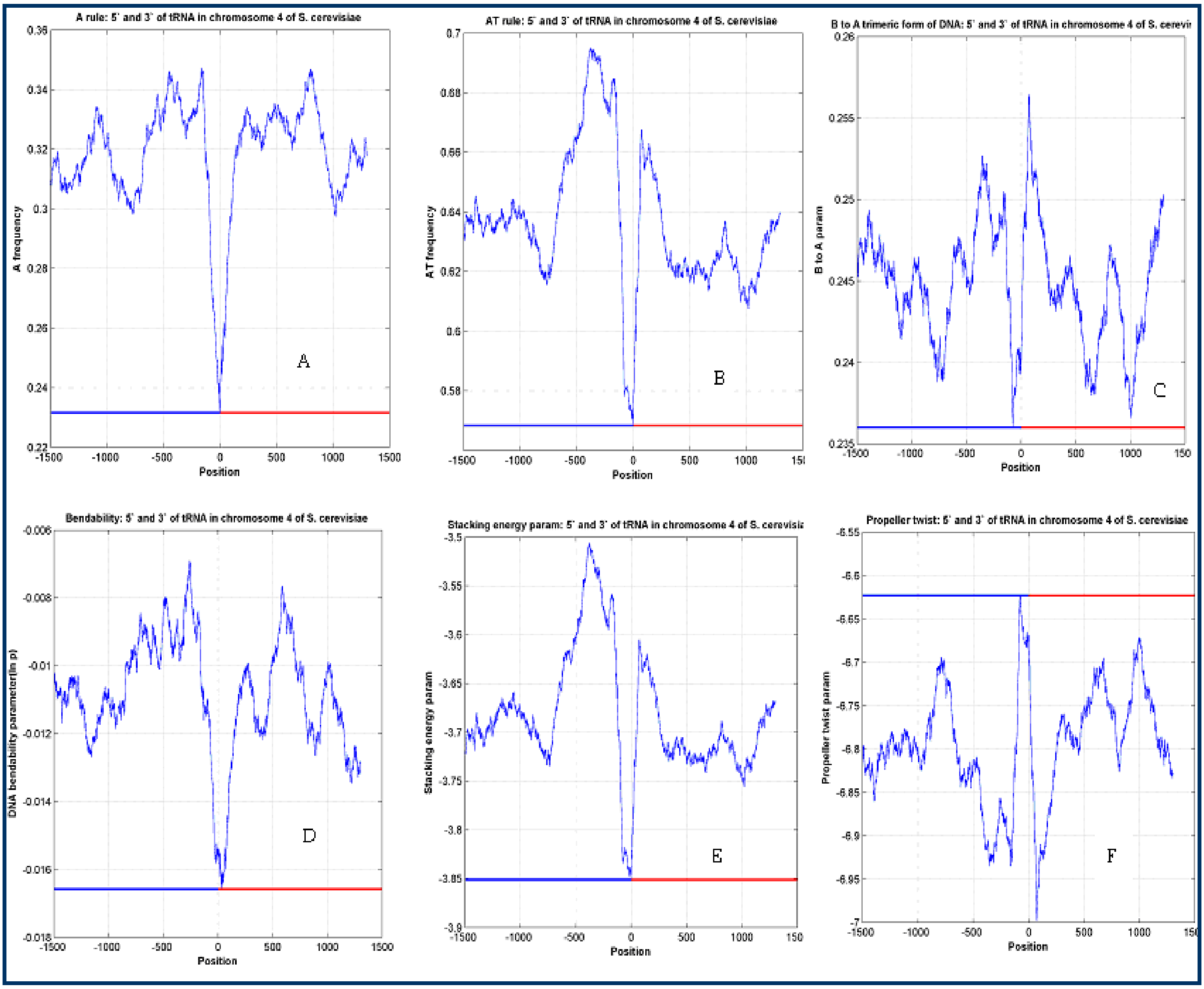

Figure 7: Various signals of the insertion sites of Ty3, for **(A, B)** A rule, AT rule **(C, D)** B to A trimeric and Bendability, **(E, F)** Stacking Energy and propeller twist. The y-axis represents value of the property and the x-axis gives the relative position with respect to the insertion site (taken to be 0). Red line indicates 5` region upstream of tRNA and blue indicates 3` downstream of tRNA.

Conclusion:

This work combines previously reported experimental work with in-silico approach to identify insertion sites, patterns, signals and endonuclease nicking sites. Our work shows that insertion sites for DNA transposon and pre-insertion site for retrotransposons have clear signals that can be useful for identifying potential insertion sites. These signals also show presence of motifs which are traditionally identified through experimental systems. This approach can help experimentalist to deduce potential motifs and can help in setting up directed experiments to find statistically and biologically important motifs at insertion sites.

References:

[1] Jurka J, *Proc Nati Acad Sci U S A*. 1997 **94**: 1872 [PMID: 9050872]
[2] Feng Q *et al. Cell*. 1996 **87**: 905 [PMID: 8945517]
[3] Liao GC *et al. Proc Nati Acad Sci U S A*. 2000 **97**: 3347 [PMID: 10716700]
[4] Mandal PK *et al. Nucleic Acids Res*. 2006 **34**: 5752 [PMID: 17040894]
[5] Rawal K & Ramaswamy R, *Nucleic Acids Research*. 2011 **39**: 6864 [PMID: 21609951]
[6] Chan PP & Lowe TM, *Nucleic Acids Research*. 2009 **37**: D93 [PMID: 18984615]
[7] Jurka J *et al. Cytogenet Genome Res*. 2005 **110**: 462 [PMID: 16093699]
[8] Bakre AA *et al. Exp ParasitoL*. 2005 **110**: 207 [PMID: 15955314]
[9] Kuduvalli PN *et al. EMBO J*. 2001 **20**: 924 [PMID: 11179236]
[10] Clark DJ *et al. J Biol Chem*. 1988 **263**: 1413 [PMID: 2447089]
[11] http://www.ecogene.org/index.php
[12] Linheiro RS & Bergman CM, *Nucleic Acids Res*. 2008 **36**: 6199 [PMID: 18829720]
[13] van Helden J *et al. J Mol Biol*. 1998 **281**: 827 [PMID: 9719638]
[14] Batzer MA & Deininger PL, *Nat Rev Genet*. 2002 **3**: 370 [PMID: 11988762]
[15] Waterston RH *et al. Nature*. 2002 **420**: 520 [PMID: 12466850]

[16] Wall JD *et al. Appl Environ Microbiol*. 1996 **62**: 3762 [PMID: 8837431]

[17] Chalker DL & Sandmeyer SB, *Genes Dev*. 1992 **6**: 117 [PMID: 1309715]

[18] Cost GJ *et al. Nucleic Acids Res*. 2001 **29**: 573 [PMID: 11139628]

[19] Parks AR & Peters JE, *J Bacteriol*. 2007 **189**: 2170 [PMID: 17194796]

Edited by P Kangueane

Citation: Dev *et al.* Bioinformation 8(16): 777-786 (2012)

Supplementary material:

Table 1: Distribution of insertion sites from EleFinder **[5]**

Element	Total elements found	Full length	5′ truncated	3′ truncated	Both ends truncated
Alu in **Human** Chromosome 22	24135	3863	4235	6249	9788
Alu in Human Chromosome X	53106	9076	11885	13062	19083
Alu in Human Chromosome Y	10121	1140	2313	1912	4756
B1 in **Mouse** Chromosome 16	7647	3833	2390	530	894
B1 in Mouse Chromosome X	11889	5791	3977	803	1318
B1 in Mouse Chromosome Y	2480	553	1227	266	434

Table 2: Comparative results of Chromosome 22 and Y and previous work **[1]** in distance class of 10-20 base pairs upstream. Ten highest occurring motifs in previous work and current work are compared

Motif	22	Y	Previous Work
TTAAAA	13.76	11.73	13.25
TTAAGA	3.81	4.59	9.5
TTAGAA	2.65	3.06	8
TTGAAA	3.20	2.55	4.25
TTAAAG	2.17	2.55	2.75
CTAAAA	1.83	0	2.5
TCAAGA	1.49	0.76	2.25
AAAAAA	7.69	6.12	2
TTTAAA	7.62	6.12	1.75
TAAAAA	12.60	10.96	1.5
GTAAGA	0.74	0.51	1.5

Table 3: Information derived from DNASCANNER analysis of the alu and B1 elements

	Alu- Human chromosome 22			B1- Mouse chromosome 16		
PROPERTIES	Trend	Position	Value	Trend	Position	Value
A rule	U	-10	0.484	U	-15	0.502
AT rule	U	-11	0.730	U	-17	0.762
B to A trimeric form	U	-10	0.277	U	-17	0.281
Bendability	D	-9	-0.021	D	-17	-0.021
Bending stiffness	D	-12	24.012	D	-17	22.526
C rule	D	-10	0.110	D	-16	0.098
Duplex stability-free energy	U	-12	-0.683	U	-17	-0.662
G rule	D	-13	0.153	D	-19	0.136
Nucleosomal positioning	D	-11	-3.371	D	-18	-3.716
Propeller twist	D	-11	-7.325	D	-15	-7.387
Protein induced deformability	D	-12	1.994	D	-18	1.957
Stabilizing energy Z-DNA	U	-12	1.776	U	-18	1.810
Stacking energy	U	-11	-3.397	U	-17	-3.321
T rule	U	-20	0.288	U	-50	0.330

Table 4: Elements analyzed and the current status of work done on them elsewhere

Repeat Element	Organism	Chromosome	Most frequent motif	References	Experimentally Proven
Alu	H. sapiens	22	TTAAAA	1	Yes
	H. sapiens	X	TTAAAA	1	Yes
	H. sapiens	Y	TTAAAA	1	Yes
B1	M. musculus	16	TTAAAA		No
	M. musculus	X	TTAAAA		No
	M. musculus	Y	TAAAAA		No
Tn7	*E. coli*	Genome	Signals found downstream of glmS	9	Yes
	Desulfovibrio desulfuricans G20	Genome	Signals found downstream of glmS	16	Yes
P-element	D. melanagoster	Insertion sites	CTCTCTCT	12	Yes
Ty3	*S. cerevisiae*	4	GTTCGA	17	Yes
	S. cerevisiae	4	GGTTCGA	17	Yes
	S. cerevisiae	4	ATTTTTTT	17	Yes
	S. cerevisiae	7	TTTTTT	17	Yes
	S. cerevisiae	7	GGTTCGA	17	Yes
	S. cerevisiae	7	GGTTCGAT	17	Yes
EhSine1	*E. histolytica*	Insertion sites	AAGGT	4	Yes