USE OF BIOINFORMATICS TO INVESTIGATE AND ANALYZE TRANSPOSABLE ELEMENT INSERTIONS IN THE GENOMES OF *CAENORHABDITIS ELEGANS* AND *DROSOPHILA MELANOGASTER*, AND INTO THE TARGET PLASMID PGDV1

A Thesis

by

ANDREA MARIAN JULIAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2003

Major Subject: Biomedical Engineering

USE OF BIOINFORMATICS TO INVESTIGATE AND ANALYZE TRANSPOSABLE ELEMENT INSERTIONS IN THE GENOMES OF *CAENORHABDITIS ELEGANS* AND *DROSOPHILA MELANOGASTER*, AND INTO THE TARGET PLASMID PGDV1

A Thesis

by

ANDREA MARIAN JULIAN

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

Gerard L. Cote
(Co-Chair of Committee)

Craig J. Coates
(Co-Chair of Committee)

Hsin-i Wu
(Member)

William Hyman
(Head of Department)

December 2003

Major Subject: Biomedical Engineering

ABSTRACT

Use of Bioinformatics to Investigate and Analyze Transposable Element Insertions in the Genomes of *Caenorhabditis elegans* and *Drosophila melanogaster,* and into the Target Plasmid pGDV1.

(December 2003)

Andrea Marian Julian, B. E., Madras University, Chennai, India

Co-Chairs of Advisory Committee: Dr. Gerard L. Cote
Dr. Craig J. Coates

Transposable elements (TEs) are utilized for the creation of a wide range of transgenic organisms. However, in some systems, this technique is not very efficient due to low transposition frequencies and integration into unstable or transcriptionally inactive genomic regions. One approach to ameliorate this problem is to increase knowledge of how transposons move and where they integrate into target genomes. Most transposons do not insert randomly into their host genome, with class II TEs utilizing target sequences of between 2 – 8 bp in length, which are duplicated upon insertion. Furthermore, amongst insertion sites, certain sites are preferred for insertion and hence are classified as hot spots, while others not targeted by TEs are referred to as cold spots.

The hypothesis tested in this analysis is that in addition to the primary consensus target sequence, secondary and tertiary DNA structures have a significant influence on TE target site preference. Bioinformatics was used to predict and analyze the structure of the flanking DNA around known insertion sites and cold spots for various TEs, to understand why insertion sites are used preferentially to cold spots for element integration. Hidden Markov Models were modeled and trained to analyze datasets of insertions of the *P* element in the *Drosophila melanogaster* genome, the *Tc1* element in the *Caenorhabditis elegans* genome, and insertions of the *Mos1*, *piggyBac* and *Hermes* transposons into the target plasmid pGDV1**.**

Analysis of the DNA structural profiles of the insertion sites for the *P* element and *Hermes* transposons revealed that both transposons targeted regions of DNA with a relatively high degree of bendability/flexibility at the insertion site. However, similar trends were not observed for the *Tc1*, *Mos1* or *piggyBac* transposons. Hence, it is believed that the secondary structural features of DNA can contribute to target site preference for some, but not all transposable elements.

DEDICATION

This thesis is dedicated to my parents Mr. Ralph Julian and Mrs. Sandra Julian who have encouraged me and provided great motivation throughout the course of this research.

ACKNOWLEDGEMENTS

I would like to thank my parents for their inspiration and support, both financial and emotional. I am extremely grateful to my co-chair Dr. Craig Coates who willingly accepted the challenge to advise me on an inter-disciplinary field of research and for being infinitely patient in teaching and helping me understand the fundamentals of molecular genetics. He allowed me the freedom to make decisions and to be creative, while he always stood by me inspiring me to find effective solutions to problems faced during the course of this research. Working with Hidden Markov Models was new to both of us and I would like to note his willingness to learn along with me and to discuss and share his views with me. I would also like to thank my co-chair Dr. Gerard Cote for encouraging me to diversify and to use engineering concepts and techniques as tools to aid research in other disciplines and for helping me better understand and validate my predictions using Hidden Markov Models. I would like to acknowledge and thank Dr. Sze for going out of his way to help me understand my model and for ensuring that I was on the right track. I specially thank Vijay Balasubramanian for helping with the JAVA programming codes used in creating all the datasets in this research. In the lab, I would like to thank all my colleagues for their support and a special word of thanks goes to Christine Gray and Paul Barron who have never hesitated to discuss and help troubleshoot research issues. I would also like to acknowledge Jeremy Brownlie for his initial ideas into this research. Finally, I would like to say a big thank you to my friend Ashwin Balasubramanian for standing by me and for urging me on when I stumbled upon problems and was completely lost.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER I : INTRODUCTION 1

CHAPTER II : BACKGROUND 3

- 2.1 Transposable elements (TEs) and transposition 3
  - 2.1.1 What are transposable elements? 3
  - 2.1.2 Discovery of transposable elements and their uses 3
  - 2.1.3 Types of transposable elements and mechanisms of transposition 4
- 2.2 Factors contributing to target site choice 8
  - 2.2.1 *P* element of *Drosophila melanogaster* 9
  - 2.2.2 *Tc/Mariner* family of transposable elements 10
  - 2.2.3 *Hermes* element of Musca domestica 12
  - 2.2.4 *piggyBac* element of Trichoplusia ni 12

CHAPTER III : BIOINFORMATICS SOFTWARE 13

- 3.1 Bioinformatics and its applications 13
- 3.2 HMMpro software 14
  - 3.2.1 What is a Hidden Markov Model? 14
  - 3.2.2 HMM Architectures 15
  - 3.2.3 Training a HMM 16
  - 3.2.4 Applications of HMM to structural analysis and pattern discovery 17
- 3.3 Weblogo software 18

CHAPTER IV : DNA PROFILE PARAMETERS 20

- 4.1 DNA bendability 20
- 4.2 Signed and unsigned nucleosome positioning 22

Page

4.3 Propeller twist 24
4.4 Stacking energy..24

CHAPTER V : MATERIALS AND METHODS...27

5.1 Software 27
5.1.1 HMMPro 27
5.1.2 Weblogo 28
5.2 Datasets 28
5.2.1 *P* element insertions in the *Drosophila melanogaster* genome 29
5.2.2 *Tc1* element insertions in the *Caenorhabditis elegans* genome..................30
5.2.3 *Hermes* element insertions in the pGDV1 target plasmid...........................31
5.2.4 *Mos1* element insertions in the pGDV1 target plasmid 33
5.2.5 *piggyBac* element insertions in the pGDV1 target plasmid........................34
5.2.6 Creating a random dataset for the pGDV1 sequences 34

CHAPTER VI : RESULTS...35

6.1 Profiles of *P* element insertions in the *Drosophila melanogaster* genome........36
6.2 Profiles of *Tc1* element insertions in the *Caenorhabditis elegans* genome 45
6.3 Profiles of *Hermes* element insertion sites in the pGDV1 target plasmid 59
6.4 Profiles of *piggyBac* element insertions in the pGDV1 target plasmid 68
6.5 Profiles of *Mos1* element insertion sequences in the pGDV1 target plasmid....78

CHAPTER VII : DISCUSSION AND CONCLUSIONS...86

7.1 Analysis of *P* element insertions in the *Drosophila melanogaster* genome 86
7.2 Analysis of the *Tc1* element insertions in the *C. elegans* genome.....................88
7.3 Analysis of the *Hermes, Mos1* and *piggyBac* element insertions into the pGDV1 target plasmid 89
7.4 Conclusions...95

REFERENCES...98

APPENDIX A 105

APPENDIX B 108

VITA 111

LIST OF TABLES

Page

Table IV-1: DNA bendability parameters as revealed by DNase I binding given as a trinucleotide scale. 21

Table IV-2: Table gives the trinucleotide signed nucleosome positioning values. 23

Table IV-3 : Table gives the dinucleotide propeller twist values.......... 25

Table IV-4 : Table gives dinucleotide stacking energy values 26

Table VI-1 : List of all transposable elements used in this study along with information on the source and integration target for each element.......... 35

LIST OF FIGURES

Page

Figure II-1: Mechanism of Class II and Class I transposition.......................................6

Figure II-2: Cut and paste mechanism of Class II transposition using inverted terminal repeats (ITRs) and the transposase protein. 7

Figure III-1: Backbone structure of the linear architecture of a HMM..........................16

Figure VI-1: DNA bendability profile of *P* element insertion sequences in the *Drosophila melanogaster* genome...37

Figure VI-2: DNA nucleosome positioning profile of *P* element insertion sequences from *Drosophila melanogaster* 38

Figure VI-3: DNA unsigned nucleosome positioning profile of *P* element insertion sequences from *Drosophila melanogaster* 40

Figure VI-4: DNA propeller twist profile of *P* element insertion sequences from *Drosophila melanogaster* 41

Figure VI-5: Stacking energy profile of *P* element insertion sequences from *Drosophila melanogaster* 43

Figure VI-6: Sequence logo of 795 *P* element insertion sites generated by the Weblogo program..44

Figure VI-7: DNA bendability profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome...46

Figure VI-8: Comparison of DNA bendability profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome...47

Figure VI-9: DNA nucleosome positioning profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome...48

Figure VI-10: Comparison of DNA nucleosome positioning profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome 49

Figure VI-11: DNA unsigned nucleosome positioning profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome 50

Page

Figure VI-12: Comparison of DNA unsigned nucleosome positioning profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome..........51

Figure VI-13: DNA propeller twist profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome 52

Figure VI-14. Comparison of propeller twist profiles of Tc1 element insertion sites in the *Caenorhabditis elegans* genome 53

Figure VI-15: DNA stacking energy profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome 54

Figure VI-16: Comparison of stacking energy profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome 55

Figure VI-17: Sequence logo of 196 *Tc1* element insertion sites generated by the Weblogo program...57

Figure VI-18: Sequence logo of the second set of 22 Tc1 element insertion sites generated by the Weblogo program 58

Figure VI-19: DNA bendability profile for *Hermes* element insertion sequences in the pGDV1 target plasmid...60

Figure VI-20: Nucleosome positioning profile of *Hermes* element insertion sequences...61

Figure VI-21: Unsigned nucleosome positioning profile of *Hermes* element insertion sequences...62

Figure VI-22: Propeller twist profile of *Hermes* element insertion sequences................63

Figure VI-23: Stacking energy profile of *Hermes* element insertion sequences.............64

Figure VI-24: Sequence logo providing the information content of the insertion sites in pGDV1 for the *Hermes* element...66

Figure VI-25: Sequence logo providing information content of the cold spots in pGDV1 for the *Hermes* element..67

Figure VI-26: DNA bendability profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid..69

Page

Figure VI-27: DNA nucleosome positioning profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid 70

Figure VI-28: DNA unsigned nucleosome positioning profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid 71

Figure VI-29: DNA propeller twist for *piggyBac* element insertion sequences in the pGDV1 target plasmid..73

Figure VI-30: DNA stacking energy for *piggyBac* element insertion sequences in the pGDV1 target plasmid..74

Figure VI-31: Sequence logo providing information content of the insertion sites in pGDV1 for the *piggyBac* element...76

Figure VI-32: Sequence logo providing information content of the cold spots in pGDV1 for the *piggyBac* element. 77

Figure VI-33: DNA bendability for *Mos1* element insertion sequences in the pGDV1 target plasmid..79

Figure VI-34: DNA nucleosome positioning for *Mos1* element insertion sequences in the pGDV1 target plasmid...80

Figure VI-35: DNA unsigned nucleosome positioning profile for *Mos1* element insertion sequences in the pGDV1 target plasmid 81

Figure VI-36: DNA propeller twist profile for *Mos1* element insertion sequences in the pGDV1 target plasmid...82

Figure VI-37: DNA stacking energy profile for *Mos1* element insertion sequences in the pGDV1 target plasmid...83

Figure VI-38: Sequence logo providing information content of the insertion sites in pGDV1 for the *Mos1* element 84

Figure VI-39: Sequence logo providing information content of the cold spots in pGDV1 for the *Mos1* element 85

CHAPTER I

INTRODUCTION

Class II transposable elements (TEs) or transposons are mobile segments of DNA that are capable of being excised and transposed from one chromosomal location to another by a process known as transposition. When a TE moves from one place to another, it can cause changes in the DNA at both the original and the target site, hence generating gene mutations and chromosomal rearrangements. TE insertions near or within a gene sequence can activate or inactivate the gene, thereby affecting gene expression. Hence, TEs are also sometimes known as 'jumping genes' and are valuable molecular tools for creating transgenic or Genetically Modified Organisms (GMOs), as well as now being implicated as playing significant roles in the evolution, structure and function of genomes. However, in some systems, the use of TEs to generate GMOs is not very efficient due to low rates of transposition and/or integration into unstable or transcriptionally inactive genomic regions. A potential solution to the latter problem lies in understanding how and where TEs integrate into the target genome so as to engineer TE movement into favorable and specific target sites.

TE movement is not completely random and exhibits variable specificity in the selection of target sites while integrating into host genomes.[1] Choice of target site can depend on several factors such as the primary sequence, transcription, replication and accessibility to chromatin[2, 3] Class II TEs use 2-8bp target recognition sequences, which are then duplicated upon insertion. However, it is clear that not all potential target sites

---

This thesis follows the style and format of the *Journal of Molecular Biology*.

within a genome are chosen with unbiased frequency. [2 - 4] Sites that have been target by TEs are called **insertion sites** and of these, those that have been used more than twice are called **hot spots**, whereas, those sites that match the primary target recognition sequence and yet remain unused are called **cold spots**. Based on the fact that TEs do not insert at every site that matches the primary target sequence, it is possible that there are other factors contributing to target choice. Hence, it is postulated that secondary and tertiary local DNA structures, such as supercoiled DNA, bending of target DNA and curved flanking DNA, may also contribute to target site preferences.

In order to better understand this concept of selective choice of TE target sites the following research objectives were proposed for this thesis:

Specific Aim 1: To use Bioinformatics as a tool to investigate the secondary structure of flanking DNA around insertion sites as well as cold spots and random DNA sequences.

Specific Aim 2: To create datasets of insertion sites of the *P* element in the *Drosophila melanogaster* genome, *Tc1* in the *Caenorhabditis elegans* genome and also undertake an analysis of insertion sites and cold spots for the *Hermes*, *Mos1* and *piggyBac* transposons into the pGDV1 target plasmid.

Specific Aim 3: To create suitable models in order to train large datasets of TE insertion sites and cold spots using Hidden Markov Model machine learning techniques in order to analyze local DNA structure.

Specific Aim 4: To choose suitable DNA profile parameters such as DNA bendability, nucleosome positioning, unsigned nucleosome positioning, propeller twist and stacking energy to predict and analyze target DNA structure.

CHAPTER II

BACKGROUND

2.1 Transposable elements (TEs) and transposition

2.1.1 What are transposable elements?

Genomes continuously evolve either by modification and mutation of existing genetic material or by the addition of new genetic material. Transposable elements (TEs) or transposons are discrete mobile sequences in the genome that can transpose themselves from one location to other locations in the genome and hence act as carriers of new genetic material.[5] A variety of names have been used to describe these genetic elements including controlling elements, cassettes, jumping genes, roving genes, mobile genes, mobile genetic elements, and transposons.[6] TEs are a heterogeneous class of genetic elements that vary in structure, mechanism of transposition and choice of target sites. They have been detected genetically through the abnormalities that they produce in the activities and structures of the genes near the sites to which they move.[6] Transposable elements of some form are found in all prokaryotic and eukaryotic organisms and include phages, bacteria, fungi, higher plants, viruses, and insects.

2.1.2 Discovery of transposable elements and their uses

The first transposons were identified in maize by Barbara McClintock in the 1940s.[7] She proved that genes could move and defined the concept of mobile genetic elements. She also found that they were responsible for a variety of gene mutations in maize. Later, *P* elements belonging to Class II TEs were found in the fruit fly, *Drosophila melanogaster*.[7] *P* elements have proved very useful in creating transgenic flies since any desired gene can be integrated into the gene by injecting the early embryo with an engineered *P* element carrying the gene.[8] Many other transposons are being studied for their use in creating transgenic insects of agricultural and medical importance.[9] Transposable elements are key to many applications in molecular genetic research. They can be used for genetic analysis as markers to tag other genes and for mutagenesis experiments to localize and characterize genes.[8, 9] Some transposons in bacteria are known to carry or mobilize genes that confer antibiotic resistance and hence impact public health.[9] However, one of the most important features of TEs is their contributions to the evolution of genomes by causing mutations as a result of insertions, deletions and recombination.[8, 9]

On the other hand, TEs have also been found to be the cause of mutations responsible for some cases of human genetic diseases, including Hemophilia A and B, porphyria and Duchenne muscular dystrophy.[9] Hence there is need to study transposable elements to explore various possibilities in transgenic research and this may also serve to better understand mutations that lead to certain diseases.

2.1.3 Types of transposable elements and mechanisms of transposition

TEs were first detected in eukaryotes and are of two types: those that mobilize by a DNA only mechanism (Class II) and those that use an RNA intermediate (Class I). Shown below in Figure II-1 is an illustration of the two different mechanisms of transposition using DNA and RNA intermediates. Figure II-1 (a) illustrates Class II transposition and uses only a DNA segment that moves directly from one place to another, also referred to as ‘cut and paste’ transposition. Figure II-1 (b) depicts Class I transposition that first transcribes the DNA copy into an RNA intermediate and then uses reverse transcriptase to make a DNA copy of the RNA intermediate to insert in a new location, hence these are often referred to as retro-transposons.

Class II transposons move by a ‘cut and paste’ mechanism wherein the transposon is excised (cut) from its original location and transposed (paste) into the new location (target site). This process of transposition shown in Figure II-2 requires an enzyme called the transposase. The transposon carries its own gene that codes for the transposase which is present in the open reading frame (ORF) within the inverted terminal repeats (ITRs). The ITRs are identical sequences reading in opposite directions. The transposon uses the host cell machinery to transcribe and translate this gene to make the transposase protein which then serves to recognize the ITRs and certain target insertion sequences and then promotes insertion of the TE at the target site. The transposase first binds to both ends of the transposon consisting of the ITRs, cuts the transposon out of the donor DNA producing sticky ends and then makes a staggered cut at the target site sequence where the transposon is pasted into the target. In the process of the transposon being excised, a gap is left behind in the original site which is then filled and nicks are sealed by host cell DNA repair mechanism.

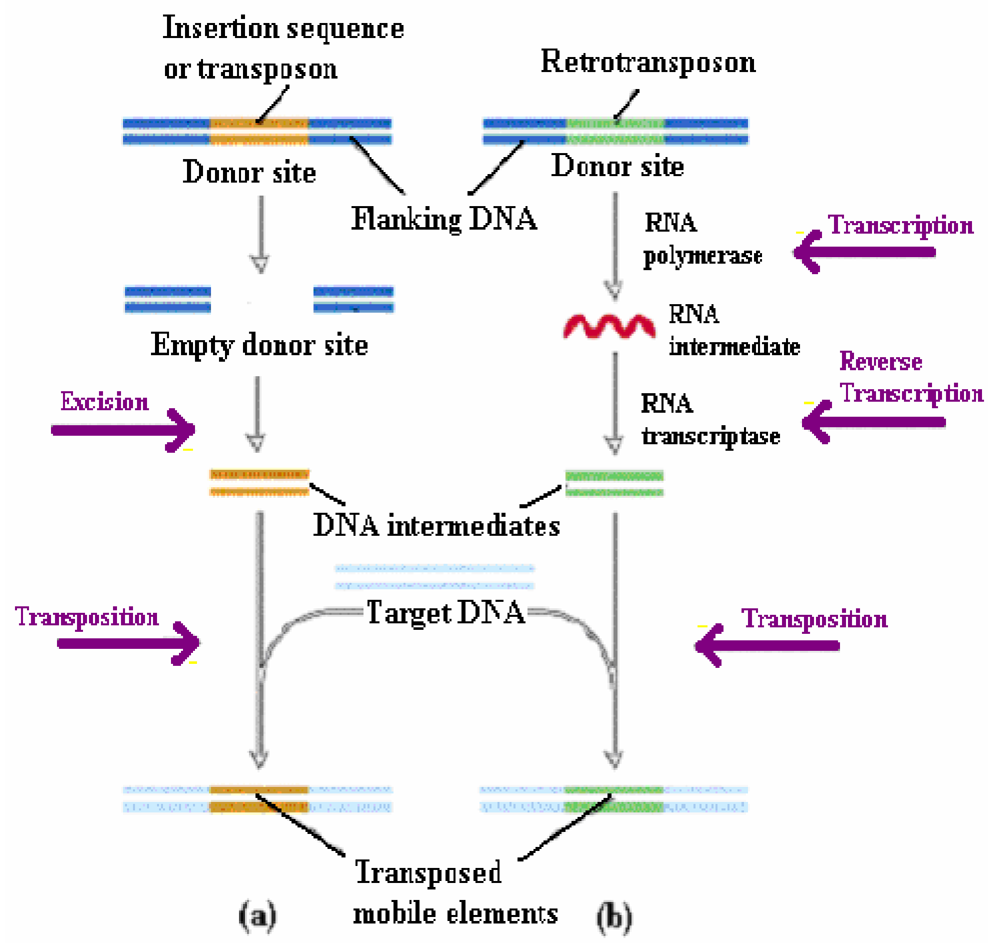


Figure II-1: Mechanism of Class II and Class I transposition. Class II transposition is shown in (a) and is a direct cut and paste mechanism whereby the transposon excises (cuts) itself out of the donor or host genome in order to transpose (paste) itself into a new location within a target genome. Class I transposition as illustrated in (b) includes transcribing the DNA into an RNA intermediate and then uses reverse transcriptase to make a DNA copy from that RNA and finally inserts the c-DNA into the target genome.

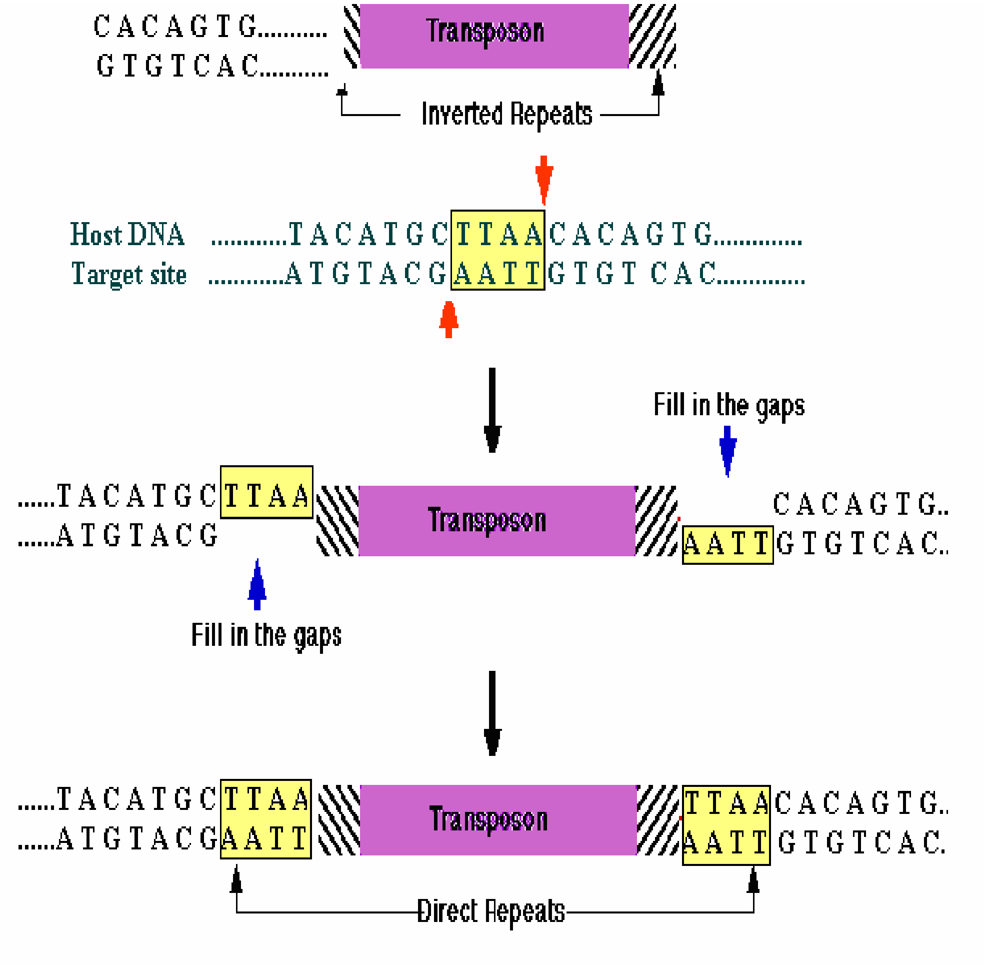


Figure II-2: Cut and paste mechanism of Class II transposition using inverted terminal repeats (ITRs) and the transposase protein. The transposase recognizes the ITRs of the transposon, cuts out the transposon from the donor site and then inserts the same into the target genome. In the process of excising the transposon from the donor site a gap is left behind which is then repaired by the host cell DNA repair machinery and further sealed.

2.2 Factors contributing to target site choice

Many research groups have studied and analyzed factors contributing to target site choice for a broad range of transposable elements.[3, 4, 10-14] Previous research shows that TEs utilize target sequences of between 2 – 8 bp in length, which is then duplicated upon insertion and yet it is also known that TEs do not hit every site that matches the primary consensus sequence. Therefore, it is hypothesized that target sequences may not be the only factor contributing to the selection of target sites. There has been an increasing interest in exploring the significance and contribution of secondary DNA structure and the potential influence of flanking sequences towards the selection of target sites.[2-4, 10]

It has been shown that in addition to the primary nucleotide sequence, certain transposable elements show a preference for particular secondary structures in target DNA, thus playing a vital role in target site selection.[15] It is believed that certain local and unusual DNA structures formed in the vicinity of hot spots enhances preferential recognition by transposition machinery and hence influences insertion of elements at those sites.[15]

In the ensuing discussion, hypothesis proposed and results established previously by various researchers with respect to target sequences and structure of flanking DNA have been addressed.

2.2.1 *P* element of *Drosophila melanogaster*

The *P* element of the fruit fly, *Drosophila melanogaster* is a small transposon with terminal 31-bp inverted repeats, and the element generates 8-bp direct repeats of target DNA sequences upon insertion. The complete element is 2907 bp and has 4 exons.[16] The *P* element is autonomous because it encodes a functional 87 kilo Dalton transposase.[17]

The *P* element demonstrates a remarkable specificity for a 8 bp GGCCAGAC consensus sequence at the target site, which is duplicated upon insertion.[16] Evidence show that *P* elements transpose non replicatively and without an RNA intermediate.[18,19] The *P* element also displays a preference for euchromatic sites over heterochromatic sites.[20] Furthermore, the *P* element exhibits a strong tendency to integrate at the 5' end of genes, in the vicinity of transcription start sites.[21-23] It was shown that target sites with close matches to the consensus octamer GGCCAGAC are more likely to receive *P* element insertions.[16, 24]

Using bendability, A-philicity, protein-induced deformability and B-DNA twist, it was also shown that DNA at the *P* element insertion sites differed significantly in structure from random DNA.[10] Flanking sequences around the *P* element insertion sites have been shown to have a high GC content and were enriched in triplets such as CAG, CTG, GAC, GCC, GGC and GTC which all have high bendability values attributed to them.[10] A graphical method called HBondView was developed to convert a set of aligned DNA sequences to a representation of potential hydrogen-bonding positions in the major groove.[10] This program indicated that *P* elements show a preference for a particular palindromic arrangement of hydrogen bonding sites over a 14 bp palindromic region centered around the insertion site.[10]

2.2.2 *Tc/Mariner* family of transposable elements

The *mariner* family[25] includes the *Tc* elements originally detected in *Caenorhabditis elegans*.[26] The *Tc1*, *Tc3* and *mariner* transposons show significant similarities. Both *Tc1* and the related *Tc3* element carry a single transposase gene, *tc1A* and *tc3A* respectively, each interrupted by one intron at a different position. The *Tc1* transposon is 1610 bp long and carries terminal inverted repeat sequences of 54 bp[27] while the *Tc3* element is 2335 bp in length with 462 bp terminal inverted repeats.[28] The autonomous copy of the *mariner* element from *Drosophila mauritiana*, *Mos1*, is 1286 bp in length and is flanked by 28 bp terminal inverted repeats.[29]

Both *Tc1* and *Tc3* transposons of *Caenorhabditis elegans* consistently integrates into TA dinucleotides which are duplicated upon insertion.[2,3,30] A 10 bp CAYATARTG consensus sequence that flanks the target TA has also been identified. [2, 3, 31 ,32] It was identified that both transposons exhibited a strong non-random preference for certain target sites that were not clustered or evenly spaced.[2] However the distribution of target sites for both elements were different and this could reflect a difference in target site choice of both elements.[2] In this regard it was hypothesized that each of these two transposons scanned sequences flanking the TA dinucleotide and recognized different patterns of preferred sequences in the close vicinity of the target site.[2] The primary sequence of the target site, as recognized by the transposase, contributes to target site choice in both *Tc1* and *Tc3* transposons of *Caenorhabditis elegans*.[2, 3, 33] Transcription and replication may also affect target site choice.[2, 3] It was shown that same type II transposons (mostly *Tc1* and *Tc3*) preferentially inserted into sites of high recombination.[34] In certain cases, it was found that the *Tc1* element had an increased affinity for supercoiled target DNA as opposed to relaxed DNA.[3] Recent research performed with *Sleeping Beauty,* a *Tc1/mariner* type transposable element in vertebrates, revealed an 8 bp palindromic AT repeat (ATATATAT) consensus sequence.[9] It was interesting that the authors found that there was indeed a difference between insertion sites and random DNA with respect to secondary structure. Several DNA structural properties were examined at the insertion sites, revealing a bendable structure in and around the target.[11]

The *mariner* element from *Drosophila mauritiana, Mos1,* integrates at TA dinucleotide residues, although it does not exhibit a strong consensus sequence around the target TA.[12, 35] The *Mos1*element also clearly did not reveal any preference for its orientation into the target pGDV1 plasmid.[12] There was also an indication for preference of GC nucleotides in the upstream and downstream flanking sequences around the insertion site.[12] The *Himar1 mariner* transposon clearly shows a bias for insertion into bent DNA target sites rather than random DNA structures.[13]

2.2.3 *Hermes* element of Musca domestica

The *Hermes* transposon is a short inverted-repeat type element that is 2749 bp in length and has 17 bp imperfect inverted repeats.[4] The *Hermes* element from the house fly, *Musca domestica,* is related to the *hobo* transposable element from *Drosophila melanogaster* and uses an 8 bp consensus target sequence, NTNNNNAC, identical to that for the *hobo* element, which is duplicated upon insertion.[4, 36-38] Chromatin structure at the target site also influences target choice for the *hobo* element.[37, 38] Certain insertion sites are most preferred for element integration with more than 2 integration events at that site and were referred to as hot spots.[4, 36] The distribution of insertion sites appear clustered around sites that serve as highly preferred integration sites and is known as the neighborhood effect.[4, 36] The *Hermes* element showed an orientation preference while integrating into the target pGDV1 plasmid and showed five times more integrations in the positive orientation than in the negative orientation.[4, 36]

2.2.4 *piggyBac* element of Trichoplusia ni

The *piggyBac* element is 2.4 kb in length and terminates in 13 bp perfect inverted repeats, with additional internal 19 bp inverted repeats located asymmetrically with respect to the ends.[39] The *piggyBac* element of *Trichoplusia ni* prefers to integrate at the tetranucleotide target sequence, TTAA, which is duplicated upon insertion.[12, 40] The *piggyBac* element shows a strong insertion site, as well as orientation preference.[12, 40]

CHAPTER III

BIOINFORMATICS SOFTWARE

3.1 Bioinformatics and its applications

Bioinformatics is a combination of Computer Science, Information Technology and Genetics to determine and analyze genetic information. Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research and can be used to gather, store, analyze, integrate and manage biological and genetic information.

Bioinformatics finds many applications. Some of the major applications involving database management include the following:

1. Creation and maintenance of databases of biological information including nucleic acid and protein sequences.
2. Storage and organization of millions of nucleotides.
3. Designing a database and developing an interface whereby researchers can both access existing information and submit new entries.

Bioinformatics is also used in more pressing tasks that involve the analysis of the integrated sequence information and this field is called computational biology. Some of the applications of bioinformatics in this field include:

1. Finding genes in the DNA sequences of various organisms.
2. Developing methods to predict the structure and/or function of newly discovered proteins and DNA/RNA sequences.

3. Clustering protein sequences into families of related sequences and the development of protein models.
4. Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

3.2 HMMpro software

HMMpro 2.2 is the newest version of Net-ID's biological sequence analysis software. Net-ID specializes in applications for computational molecular biology using machine learning techniques and object oriented software design.[41, 42] HMMpro is a biological sequence analysis package based on hidden Markov model machine learning techniques built on top of the foundation libraries. HMMpro is a general purpose HMM (Hidden Markov Model) simulator for the modeling, analysis, classification, and alignment of biological sequences. It can be used in data base searches, multiple alignments, and pattern discovery to study genes and regulatory regions, or the structure and function of protein families. Net-ID uses state-of-the-art computer technology which includes Object Oriented Programming in Java and C++.

3.2.1 What is a Hidden Markov Model?

A HMM is a stochastic generative model for time series defined by a finite set S of states, discrete alphabet A of symbols, probability transition matrix $T= (t_{ij})$ and a probability emission matrix $E= (e_{ix})$. The alphabet A includes the DNA/RNA alphabets consisting of 4 nucleotides and the protein alphabet consisting of 20 amino acids. The system randomly evolves from one state to the other while emitting symbols from the alphabet. In a given state i, it has a probability t_{ji} of moving to state j and probability e_{ix} of emitting symbol X. It is based on the assumption that emissions and transitions depend only on current state. The reason why they are referred to as ‘hidden’ is because only the symbols emitted by the model are observable and not the underlying random walks between states.[41]

3.2.2 HMM Architectures

The architecture of an HMM is the graph associated with the HMM states and the non-zero probability transitions.[41] There are various architectures such as the linear, tied, loop, wheel, parallel and hybrid types. The linear or standard architecture shown in Figure III-1 is the one most commonly used. With any architecture there are two special states namely the start state (S) and the end state (E). In addition to these two states are three other classes of states namely the main or match states (M), delete states (D) and the insert states (I). The delete states are also called gap or skip states. The main and insert states will emit alphabets of proteins/DNA/RNA while the delete states are mute. The self-transitions on the insert states are one way of allowing for multiple insertions. The sequence of states leading from the start state to a series of main states and finally to the end state is called the backbone of the model. A linear architecture is specified by giving the length of its backbone, i.e. its total number of main states.

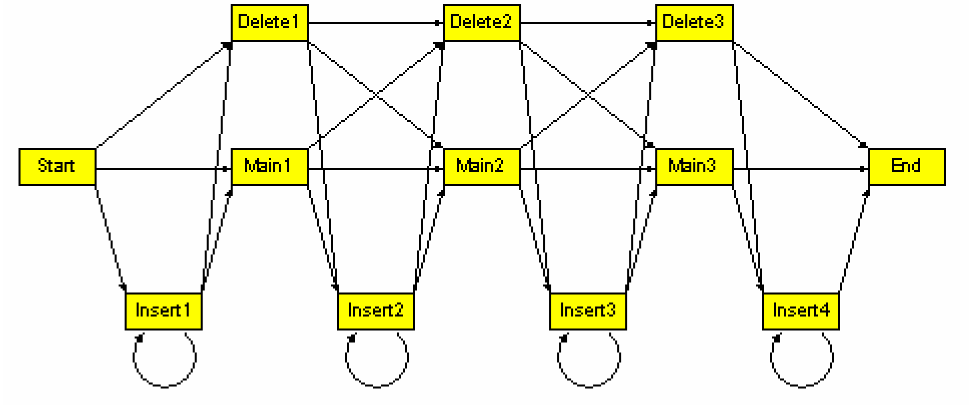


Figure III-1: Backbone structure of the linear architecture of a HMM. The backbone of this model is given by the linear sequence connecting the start state with the end state through intermediate main states. Delete states are dummy states that account for gaps while the start, main and insertion states emit alphabets.

3.2.3 Training a HMM

HMM training or learning is statistical model fitting.[41] Given a set of sequences, HMM parameters can be estimated by Maximum Likelihood (ML) or Maximum A Posteriori (MAP) estimation. In general, ML or MAP solutions cannot be derived analytically but approximated using one of several possible iterative algorithms such as gradient descent or the EM (Expectation-Maximization) [also known as Baum-Welch] algorithm. The learning rate is a parameter that governs the size of the iterative steps taken in parameter space when doing gradient descent on the negative log-likelihood.

Learning is said to be on-line, when parameters are modified after the presentation of one or a small number of training examples. It is said to be off-line or batch if parameters are modified only after the presentation of all or a large number of training examples. On-line learning is often preferable because of its flexibility with respect to data and storage, and because the element of stochasticity introduced at each step by the choice of training example can be useful in avoiding poor local optima.

In typical gradient descent or EM algorithms, iterations are based on the calculation of sequence likelihoods using the forward procedure. The exact calculation of the parameter updates requires also a symmetric dynamic programming procedure called the backward algorithm. Hence the name forward-backward algorithm for the algorithm that produces the value of the parameter updates in many of the iterative learning algorithms. The backward algorithm computes probabilities of being in each HMM state backwards in time.

The term Viterbi learning refers to any form of learning where only the optimal paths associated with the training sequences are used to determine the parameter updates, as opposed to all possible paths associated with the forward algorithm and the computation of likelihoods. In general, Viterbi learning works well with large alphabets and homologous sequences (protein families), but less so with small alphabets and non-homologous sequences (DNA exons or promoters) where full gradient descent or full EM are preferable.

3.2.4 Applications of HMM to structural analysis and pattern discovery

Information about new patterns and structure can be identified from a trained HMM.[41] High emission or transition probabilities are normally associated with conserved regions or consensus patterns that may have structural or functional significance. One method would be to plot the entropy of the emission distributions along the backbone of the model and the other method would be to use features such as protein hydrophobicity or DNA bendability which can then be averaged and plotted using the HMM probabilities. Patterns characteristic to a particular class or family such as secondary structural features are easier to detect in the HMM plots.

3.3 Weblogo software

Positional dependent information of contents of aligned RNA/DNA or amino acid sequences are useful for the display of consensus sequences and for finding optimal search windows used in sequence analysis. The simplest form of a consensus sequence is created by picking the most frequent base at some position in a set of aligned DNA, RNA or protein sequence. The process of creating a consensus destroys the frequency information and leads to many errors in interpreting sequences. If a position at a site had a 75% occurrence of the 'A' nucleotide, then the consensus would be 'A'. Hence, distinction between 100% A and 75% A cannot be made. This approximate estimation of data at a position leads to wrong predictions of genetic data. This problem can be eliminated using sequence logos wherein every nucleotide is represented according to its frequency of appearance. Subtle frequencies are not lost in the final product as they would be in a consensus sequence. The sequence logo shows not only the original frequencies of the bases, but also shows the conservation at each position and since it is graphic, patterns exhibited in a profile are immediately revealed.

Weblogo[43] is a web based program used in the generation of sequence logos. Sequence logos are a well known way to visualize a profile or a multiple alignment. A sequence logo is a graphical display that provides information about the frequencies of bases at each position, as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters.[44, 45] The X-axis is a representation of position while the vertical scale is in bits, with a maximum of 2 bits possible at each position. [44, 45] The height of each letter is drawn proportional to its frequency and the letters are sorted so that the most frequent one is on top. Sequence conservation at a position is measured in bits of information. The binary digit ‘bit’ is the choice between two equally likely possibilities. There are 4 bases in DNA, and these can be arranged in a square:

A C
G T

To pick one of the 4 it is sufficient to answer only two yes-no questions: ‘is it on top?’ and ‘is it on the left?’. Thus the scale for the sequence logo runs from 0 to 2 bits. For 8 parameters it takes 3 bits and so on and so forth. When the frequencies of the bases are not exactly 0, 50 or 100 percent, more sophisticated calculations must be made.

CHAPTER IV

DNA PROFILE PARAMETERS

4.1 DNA bendability

The trinucleotide bendability model of Brukner et al.[46] was based on DNase I cutting frequencies. These experimentally determined trinucleotide values are reflective of the anisotropic flexibility or bendability of a particular DNA sequence. A DNA binding protein such as DNase I is considered a good molecular probe of bendability since DNase I preferentially binds and interacts with a 6 bp surface on the minor groove and cuts DNA that is bent or bendable away from the enzyme and towards the major groove (positive roll).[47, 48] Hence it is believed that DNA that is more bendable would be more accessible to DNase I cleavage.[46] Thus DNase I cutting frequencies is a direct measure of major groove compressibility or anisotropic flexibility. The bendability scale corresponds to 32 complementary trinucleotides that range from -0.28 (rigid) to +0.194 (bendable). Table IV-1 below reveals the DNase I - derived trinucleotide bendability scales.

Various other bendability models have also been proposed but the two most popularly used trinucleotide models are the DNase I – derived bendability model and the nucleosome positioning model.[49]

DNase I based trinucleotide bendability scale

Trinucleotide step	DNase-I based trinucleotide value
AAT/ATT	-0.280
AAA/TTT	-0.274
CCA/TGG	-0.246
AAC/GTT	-0.205
ACT/AGT	-0.183
CCG/CGG	-0.136
ATC/GAT	-0.110
AAG/CTT	-0.081
CGC/GCG	-0.077
AGG/CCT	-0.057
GAA/TTC	-0.037
ACG/CGT	-0.033
ACC/GGT	-0.032
GAC/GTC	-0.013
CCC/GGG	-0.012
ACA/TGT	-0.006
CGA/TCG	-0.003
GGA/TCC	0.013
CAA/TTG	0.015
AGC/GCT	0.017
GTA/TAC	0.025
AGA/TCT	0.027
CTC/GAG	0.031
CAC/GTG	0.040
TAA/TTA	0.068
GCA/TGC	0.076
CTA/TAG	0.090
GCC/GGC	0.107
ATG/CAT	0.134
CAG/CTG	0.175
ATA/TAT	0.182
TCA/TGA	0.194

Table IV-1: DNA bendability parameters as revealed by DNase I binding given as a trinucleotide scale. High values indicate bending towards the major groove. The more positive the value or the closer the value is to a zero, the more bendable the DNA is in that region whereas, the more negative or further away the value is from zero, the more rigid the structure.

4.2 Signed and unsigned nucleosome positioning

The DNA nucleosome positioning scale is also a triplet scale. Experimental investigations of DNA positioning on nucleosomes have revealed that certain nucleotides have a strong preference for being positioned with their minor grooves facing either towards or away from the nucleosome core.[50-52] The positioning of trinucleotides in helices wrapped around nucleosomes was studied and it was found that this was determined by their bending propensity towards the major groove.[50, 52] Satchwell et al.[50] determined the occurrence of individual trinucleotides facing towards (facing in) and away (facing out) from the nucleosome core. This was further scaled to roll angles by Goodsell et al.[51]

Hence, position preference is a measure of helix flexibility based on a set of 32 trinucleotide values giving the log-odds of the minor groove facing outwards when wrapped around the nucleosome core. In this model, all triplets with a close to zero position preference are considered flexible, whereas, those triplets with large absolute values that have a preference for position and may face either in or out are considered rigid. A measure of flexibility is obtained by removing the sign from the original trinucleotide values giving rise to absolute or unsigned nucleosome positioning preference.[53, 54] Table IV-2 gives the trinucleotide signed nucleosome positioning values.

Trinucleotide signed nucleosome positioning parameter scale

Trinucleotide step	Signed nucleosome positioning trinucleotide value
GCC/GGC	+45
TCG/CGA	+31
AGC/GCT	+25
CGC/GCG	+25
CAT/ATG	+18
CAC/GTG	+17
GGG/CCC	+13
TGC/GCA	+13
AGT/ACT	+11
GAG/CTC	+8
GGT/ACC	+8
TGG/CCA	+8
CGT/ACG	+8
AGG/CCT	+8
GAC/GTC	+8
TGA/TCA	+8
GAT/ATC	+7
TGT/ACA	+6
AAG/CTT	+6
CGG/CCG	+2
CAG/CTG	-2
GGA/TCC	-5
TAC/GTA	-6
AAC/GTT	-6
AGA/TCT	-9
CAA/TTG	-9
GAA/TTC	-12
TAT/ATA	-13
TAG/CTA	-18
TAA/TTA	-20
AAT/ATT	-30
AAA/TTT	-36

Table IV-2: Table gives the trinucleotide signed nucleosome positioning values. This scale can also be used as an absolute scale wherein as values get closer to zero, the less preference for specific positions in the nucleosome and hence flexibility is inferred. Similarly, values further away from a zero would indicate rigid regions.

4.3 Propeller twist

The DNA propeller twist scale is a dinucleotide scale. Propeller twist is a measure of helix rigidity as these twist angles have been shown to be inversely related to the rigidity of DNA.[55] A correlation between the propeller twist angle in base pairs and the dinucleotide step that they represented was established.[55] It was shown that regions of DNA with higher twist angles exhibited a locking effect that made those base pairs rigid.[55] Hence, regions of high propeller twist would indicate helix rigidity in that area and similarly regions of that were quite flexible would have a low propeller twist angle. The highest propeller twist of -18.66 belongs to the AA (=TT) step suggesting rigidity while the lowest propeller twist of -8.11 belongs to the GG (=CC) step which indicates flexibility. Table IV-3 gives the dinucleotide propeller twist values.

4.4 Stacking energy

Stacking energy relates to the interaction energy between adjacent base pairs in the double helix. It is estimated using a set of dinucleotide values determined by quantum mechanical calculations on crystal structures given by Ornstein et al.[56] All stacking energies are negative because base stacking is an energetically favorable interaction that serves to stabilize the double helix. It is expressed in kcal/mol and ranges from -3.82 (will melt easily) to a maximum value of -14.59 (requires most energy to destack or melt the helix). Hence a positive peak in base stacking or values closer to zero reflects regions of the helix which would destack or melt more easily. Conversely, larger negative numbers would represent more stable regions of the DNA helix. Table IV-4 represents the dinucleotide stacking energy values.

Dinucleotide propeller twist parameter values

Dinucleotide step	Propeller twist (degrees)
AA	-18.66
AC	-13.10
AG	-14.00
AT	-15.01
CA	-9.45
CC	-8.11
CG	-10.03
CT	-14.00
GA	-13.48
GC	-11.08
GG	-8.11
GT	-13.10
TA	-11.85
TC	-13.48
TG	-9.45
TT	-18.66

Table IV-3: Table gives the dinucleotide propeller twist values. This is a scale that is directly related to rigidity. Higher twist angles (more negative values) are indicative of higher rigidity while lower twist angles (less negative values) suggest flexibility.

Dinucleotide base stacking energy parameter values

Dinucleotide step	Stacking energy (kcal/mole)
AA	-5.37
AC	-10.51
AG	-6.78
AT	-6.57
CA	-6.57
CC	-8.26
CG	-9.69
CT	-6.78
GA	-9.81
GC	-14.59
GG	-8.26
GT	-10.51
TA	-3.82
TC	-9.81
TG	-6.57
TT	-5.37

Table IV-4: Table gives dinucleotide stacking energy values. This parameter is a function of stability of the DNA. Less negative energy values refer to higher stability than the more negative energy values.

CHAPTER V

MATERIALS AND METHODS

5.1 Software

Different software packages were employed in the creation of the DNA structural profiles. HMMPro is a biological sequence analysis package that uses Hidden Markov Modeling (HMM) techniques and was used to generate DNA profiles of the different transposable elements. Information content of the different positions of the upstream and downstream flanking sequence as well as the target site was determined for the different elements using software called Weblogo.

5.1.1 HMMPro

A HMM model was initiated by creating a new model (to be trained). Options for building a model include; (a) Alphabet: DNA or Protein or Other (b) Type of Architecture (Linear, Loop, Wheel, or Parallel) (c) Model Length (in general, the average length of the sequences being modeled) (d) Connectivity (Basic or Full).For DNA applications, it is normally recommended to use a Linear Architecture with Basic Connectivity. The HMM was created setting alphabet as DNA, linear architecture and model length equal to the average length of the sequences being modeled.

To read a file of sequences into the HMM, certain appropriate formats such as the fasta or HMM internal format should be used. Hence, files containing sequences in FASTA format were used for analysis. After reading the set of sequences, the model then needs to be trained using appropriate algorithms from the training options. For

DNA models it is preferable to use the Full Gradient Descent training algorithm and hence this algorithm was coupled with an online method of training where parameter values were modified consistently. The number of training iterations across the entire training set was set to a small sufficient number between 5 and 10 iterations.

Finally an analysis of the set of sequences was performed using DNA profile parameters. The five parameters being considered in this analysis are DNA bendability, [46, 49, 54] nucleosome positioning, [50, 51] unsigned nucleosome positioning, [53] propeller twist [55] and stacking energy.[56]

5.1.2 Weblogo

The weblogo software accepts input of sequences in fasta format. In this case, the input consisted of multiple DNA sequences. The weblogo created was 25cm x 15cm in size. The starting position and range of the weblogo profile was specified. Advanced image options included bitmap resolution and a color scheme to depict each nucleotide by a different color.

Sequence logos were generated for all datasets described below. The Weblogo outputs illustrate conserved sequences at insertion sites and also provided a picture of the information content of the flanking sequences.

5.2 Datasets

In terms of analysis, large datasets were chosen such that the investigated elements had hundreds of insertions in a genome. Datasets of *P* element insertions in the

Drosophila melanogaster genome and *Tc1* element insertions in the *Caenorhabditis elegans* genome were chosen.[2, 57-59] Insertions of the *Hermes, piggyBac* and *Mos1* elements into the pGDV1 target plasmid, as revealed by plasmid based transposition assays, were also analyzed.[4, 12, 14, 35, 36, 39]

5.2.1 *P* element insertions in the *Drosophila melanogaster* genome

The Berkeley *Drosophila* Genome Project (BDGP) extensively studied *P* element insertions.[21] They undertook to understand gene function by insertional mutagenesis, in which they used an engineered *P* element called the EP element.[60, 61] Information for the *P* transposable element insertions was derived from the Berkeley Drosophila Genome Project website.[57]

The file obtained from this website contained 4218 *P* element insertions and had information on the exact insertion point, along with a variable length of flanking sequence. These sequences were readily available in FASTA format, compatible with the HMMpro software. 100 bps of flanking sequence on each side of the insertion point were examined. In order to do this, each insertion point was aligned one below the other, along with their corresponding 100 bps each of upstream and downstream flanking sequence. Hence, effectively, each sequence was 200 bps in length, centered about the insertion point. Since the original sequence file consisted of sequences with variable lengths of flanking sequence, those sequences that fit these requirements were first selected.

A java code was written to edit the dataset as follows. The header information was read, the exact location of the insertion point bp in the DNA sequence (usually obtained by inverse PCR) was identified, then 100bps upstream and downstream of the insertion point was counted and finally the desired sequence was cut and pasted into an output file that could be retrieved for later use. While doing so, it was ensured that there were no unknown alphabets in the sequence output and also that there were indeed 100 bps on either side of the insertion point. After filtering out sequences that did not meet these requirements, 795 sequences were obtained from the original dataset of 2454 sequences.

A suitable model was designed, the sequence file was read into the HMMpro software, the model was trained and final analysis of the dataset of *P* element insertions was performed using DNA profile parameters as estimates. To examine larger flanking sequences, 336 sequences were obtained that were 400 bps in length, with 200 bps of flanking sequence on either side of the insertion point. A random dataset was generated from the original 2454 sequences, without consideration of the insertion point. This dataset was obtained by aligning the initial 200 bps of each *P* element insertion sequence from the initial file of 2454 sequences.

5.2.2 *Tc1* element insertions in the *Caenorhabditis elegans* genome

In this study two different datasets of *Tc1* insertions were analyzed. Information regarding *Tc1* transposon insertion sites in *Caenorhabditis elegans* is available in detail at the *C. elegans* Genome Project website.[2, 58, 59, 61] The first dataset was obtained from a list of *Tc1* alleles resulting from a shotgun sequencing approach.[58, 59, 62]

There were a total of 821 sequences, however, they were not as readily accessible as the *P* element insertions. Each sequence had a _L or _R designation, which indicated which end of the *Tc1* element the sequence trailed off from. Each sequence also started with the TA dinucleotides that the *Tc1* element had inserted into. These sequences were then used to blast against the entire *C. elegans* genome sequence data and recovered 196 useful sequences with 200bp of flanking sequence on each side of the TA target site. Having performed this operation, these sequences were then modeled and trained using HMMpro and performed an analysis on the structure of the DNA flanking the insertion sites in terms of the DNA profile parameters.

The second dataset consisted of 22 non-repetitive independent *Tc1* insertions in a 1 Kbp region of the gpa-2 gene within the *C.elegans* genome.[2] The gpa-2 gene was located on chromosome V of the *C.elegans* genome. Based on information of the target site consensus sequences,[2] 100bps of flanking sequence on either side of the TA target site were retrieved. Similar to the first dataset of *Tc1* insertions, this dataset was also modeled, trained and analyzed.

5.2.3 *Hermes* element insertions in the pGDV1 target plasmid

The dataset for *Hermes* insertions was derived from the results established by previous research performed with the *Hermes* element, using plasmid based transposition assays.[4, 36] It was observed that there were certain striking features in the selection of target sites, such as the use of only 65 out of 3852 potential target sites that could be used, a distinctive orientation preference and neighborhood effect. [4, 36] Furthermore, the distribution of insertion sites appear clustered around sites that served as highly preferred integration sites (736+, 2154+, 2303+). Preferred *Hermes* integration sites seem particularly clustered at approximately 80bps and 160 bps 5' and 3' of the site 736+ and these preferred sites were separated by regions that served as poor targets.[4, 36] Of these 65 insertion sites, 22 sites were targeted two or more times and hence were classified as being preferred or hot spots for integration of the *Hermes* element. Three of these 22 sites in particular had 10 or more insertions. Four possible consensus sequences were identified from the 65 insertion sites, 3 of these being associated with the 22 preferred sites, with the hottest spots using a single predominant consensus. The 8 bp NTNNNNAC consensus sequence was the most widely used among the 4 consensus sequences that were identified for the *Hermes* element.[4] Eleven further insertion sites for the *Hermes* element were identified based on transposition assays performed by Sarkar et al.[36]

The dataset of *Hermes* insertions that was examined contained a total of 76 insertion sites matching 1 of the 4 consensus sequences. In the creation of the dataset of potential target sites, or cold spots, every potential target site within the pGDV1 plasmid that matched the most frequently used 8 bp NTNNNNAC consensus sequence was identified, excluding the chloramphenicol resistance gene. Consensus patterns in both + and – orientations were checked with respect to the orientation of the chloramphenicol resistance gene open reading frame. Having thus located all potential consensus target sites, sites that were previously used as insertion sites were eliminated. It was concluded that the remaining sites that matched the 8 bp consensus pattern, yet had never been targeted previously, were unused, or cold spots.

A java code was written to edit these sequences in a similar manner as described previously. 100 bps of each flanking sequence, with the consensus pattern in the middle, were aligned to form a dataset. Consequently, 76 sequences for the insertion sites and 51 sequences for the unused cold spots were retrieved. Both datasets were analyzed for comparison and hence to infer if there was any significant difference in flanking DNA structure.

5.2.4 *Mos1* element insertions in the pGDV1 target plasmid

The approach used in creating the *Mos1* element dataset was identical to that used for the *Hermes* insertions. From previous assays performed with the *Mos1* element, it is clear that unlike the *Hermes* element, *Mos1* does not show any particular preference for insertion orientation into pGDV1 and neither does it possess a conserved consensus sequence other than the target TA dinucleotides which it duplicates upon insertion.[12, 35] In a similar manner, a java code was written that identified all potential TA dinucleotides within the pGDV1 plasmid and further segregated these sites as either being an insertion site or an unused site based on prior knowledge of target sites that had been previously hit during transposition assays. 180 potential target sites were retrieved, of which 35 were previously hit insertion sites and hence the remaining 145 sites were designated as unused or cold spots.

5.2.5 *piggyBac* element insertions in the pGDV1 target plasmid

Using the known target sites for *piggyBac* insertions based on transposition assays, it was observed that very few positions in the pGDV1 plasmid are utilized.[14, 40] Unlike *Hermes* or *Mos1*, the *piggyBac* insertions resulted in duplications that were either precise or had partial deletions of the target sequence. [14, 40] The *piggyBac* element shows a strong preference for orientation and targets the tetranucleotide TTAA. [14, 40] Based upon these details, this dataset was created using the identical method employed for the *Hermes* and the *Mos1* elements, looking for TTAA residues to identify potential target sites. This dataset was the smallest used and 24 potential target sites were retrieved, which were subsequently grouped as 15 insertion sites and 9 cold spots.

5.2.6 Creating a random dataset for the pGDV1 sequences

Similar to the concept for generating the random *P* element sequences, 2268 pGDV1 sequences, each of 400 bps in length were randomly generated. A java program simplified this process by aligning each consecutive 400 bps within the pGDV1 plasmid. In short, n to (n+399) bps where n=1, 2, 3… were used and these sequences were aligned one below the other. In the process sequences within the Cam^R gene were not included and then the same operation was performed using the negative strand of pGDV1.

CHAPTER VI

RESULTS

The hypothesis presented in this research is that even when there is a consensus target site present, there are structural conformations caused by flanking DNA sequences that can either prevent or promote transposable element insertion. Bioinformatics was used as a tool to investigate DNA helix architectural parameters such as DNA bendability, nucleosome positioning, unsigned nucleosome positioning, propeller twist and stacking energy based on dinucleotide and trinucleotide scales.

Table VI-1 provides a summary of the TEs examined in this study, along with information on the source of each element and the target sequence into which it integrated.

S.No	**Transposable element**	**Source**	**Analysis performed in genome/plasmid**
1	*P* element	*Drosophila melanogaster*	*Drosophila melanogaster*
2	*Tc1* element	*Caenorhabditis elegans*	*Caenorhabditis elegans*
3	*Mos1* element	*Drosophila mauritiana*	pGDV1
4	*Hermes* element	*Musca domestica*	pGDV1
5	*piggyBac* element	*Trichoplusia ni*	pGDV1

Table VI-1: List of all transposable elements used in this study along with information on the source and integration target for each element.

6.1 Profiles of *P* element insertions in the *Drosophila melanogaster* genome

Structural profiles of *P* element insertion sites in the *Drosophila melanogaster* genome are shown. The DNA bendability profile (Figure VI-1) gives a comparison of the insertion sites, represented as the *P* flank 200, and the random DNA sequences, represented as *P* random 200. In figure VI-1 and likewise all of the following figures, the start of the 8 bp target site consensus sequence is at position 0 and accordingly base pairs upstream of the target sequence are denoted by '-'whereas, base pairs downstream are denoted by '+'.

The insertion sites for the *P* element seem to contain fairly bendable DNA in and around the insertion site for about 15 bps on either side. The average value of bendability at the middle of the 8bp insertion site consensus sequence (position +4) is -0.01616. The random DNA sequences show a lesser pattern of bendability in comparison to the integration sites and have an average value of about -0.02209 at the potential insertion site. Looking at the general trend in profile pattern before, at and after the insertion point, there seems to be a gradual transition from being bendable to relatively less bendable to most bendable and then again back to less bendable and finally bendable again.

Consistent with the bendability profile, the *P* element signed nucleosome positioning profile (Figure VI-2) also suggests that the DNA flanking the insertion sites is relatively more flexible than the random DNA. The insertion sites show a higher average value of nucleosome positioning (4.847919) than the random DNA (0.049312). Similar to the bendability profile, nucleosome positioning also shows similar trends wherein the regions immediately before and after the insertion point are regions of least flexibility, flanked by relatively flexible regions further out. It was observed that the middle of the insertion consensus sequence lies in the upslope region tending from being least flexible to most flexible.

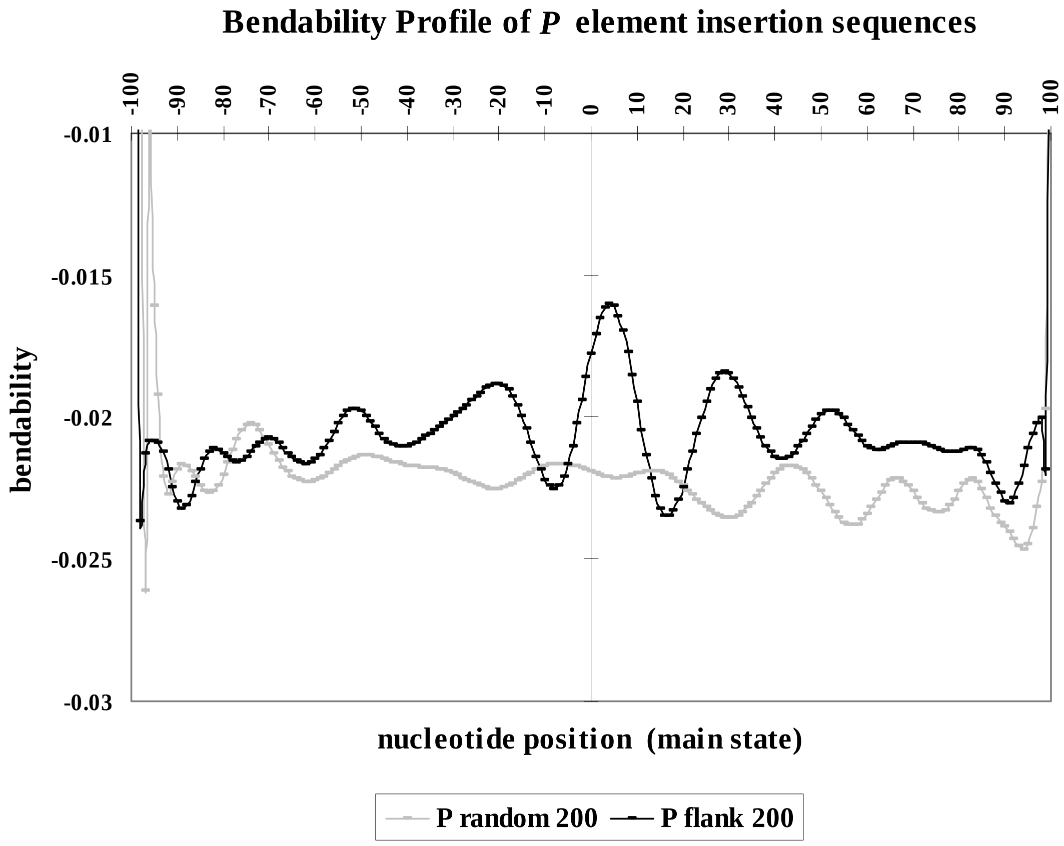


Figure VI-1: DNA bendability profile of *P* element insertion sequences in the *Drosophila melanogaster* genome. The profile represents a model that analyzed 795 insertion sites (*P* flank 200) depicted by the black line and 2454 random DNA sequences (*P* random 200) depicted as the gray line. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the bendability values. Bendability is a measure of anisotropic flexibility and is given by a trinucleotide scale. The more positive or closer the value to zero implies a greater degree of bendability, whereas more negative values indicate rigidity. As seen in this figure, there is a difference in the values of bendability between the insertion sites and the random DNA, particularly at and around the insertion point. The profile exhibits a change in trend from being less bendable before the insertion point to being most bendable at the insertion point and then again least bendable in the region after the insertion point. The bendability profile of the *P* element also exhibits a fairly symmetrical trend around the insertion site.

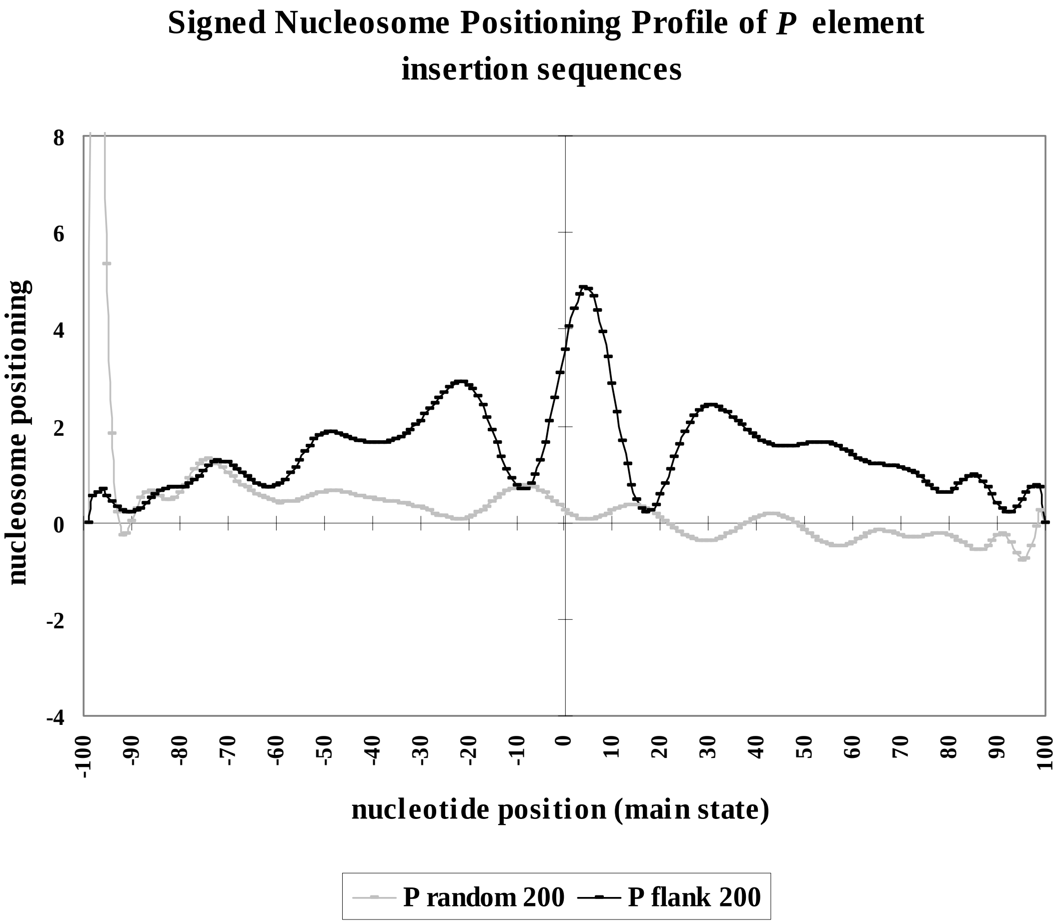


Figure VI-2: DNA nucleosome positioning profile of *P* element insertion sequences from *Drosophila melanogaster*. The profile represents a model that analyzed 795 insertion sites (*P* flank 200) depicted by the black line and 2454 random DNA sequences (*P* random 200) depicted as the gray line. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis represents the signed nucleosome positioning values. The higher or more positive the value of nucleosome positioning, the greater the flexibility of DNA in that region. A clear change in pattern is seen in the DNA immediately flanking the insertion sites. Comparable to the trends seen in the case of bendability, nucleosome positioning also indicates that there are visible transitions in patterns going from less bendable or rigid regions of DNA upstream of the insertion point to most bendable regions very close to the insertion point and then back to rigid regions of DNA downstream of the insertion point. The random DNA sequences reveal less flexibility in general.

The unsigned nucleosome positioning profile of the *P* element is shown below in Figure VI-3. This trinucleotide scale is based on absolute values and is interpreted as having zero preference to particular positions or a measure of flexibility when values are close to zero. On the other hand, values further away from a zero or more positive values would indicate rigidity. The profile in figure VI-3 also suggests differences in trends between the insertion sites and the random DNA in and around the target site. The insertion sites show a lower positive value of nucleosome positioning (+13.81271) than the random DNA (+14.0718). Hence, the DNA flanking the insertion sites is relatively more flexible than the random DNA. The insertion sites in this profile reveal a trend exhibiting regions of lesser flexibility immediately before and after the target site while the target site is in itself relatively more flexible in that region. It is seen that the middle of the insertion consensus sequence lies in the upslope region tending from being least flexible to most flexible. The opposite trend is seen with the signal from the random DNA

The propeller twist profile of the *P* element is shown in Figure VI-4. Propeller twist is a measure of DNA rigidity and these two parameters are known to be inversely related to each other. This dinucleotide scale consists of only negative values and is interpreted such that low twist or less negative values correspond to regions of flexibility whereas high twist or more negative values correspond to regions of inflexibility. As can be clearly seen in figure VI-4, the insertion sites have an average twist value of -12.3629 at the target site while the random DNA has an average twist value of -13.1038 at the target site. Comparing these two values it can be concluded that the insertion sites represented as *P* flank 200 exhibits more flexibility in the vicinity of the target site than the random DNA. As with the previous profiles, this profile corresponding to the propeller twist also reveals significant changes in trends between the insertion sites and the random DNA.

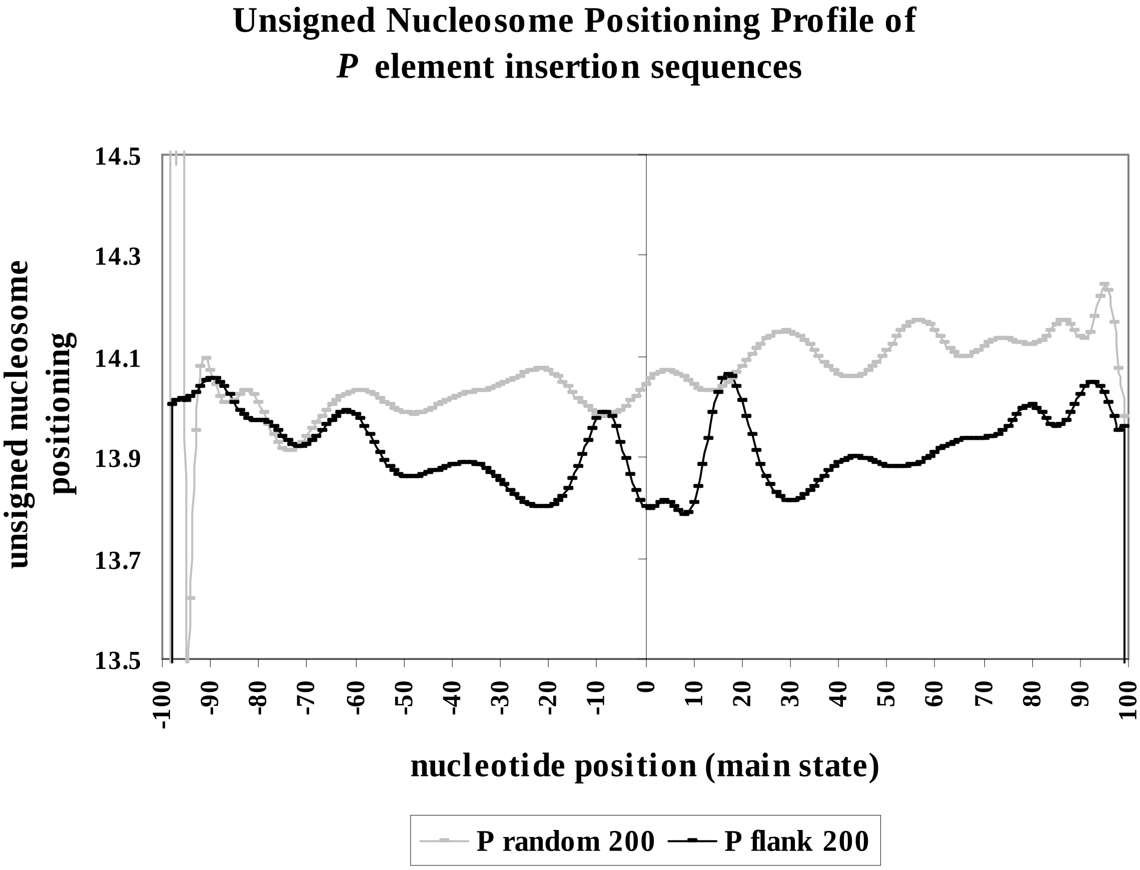


Figure VI-3: DNA unsigned nucleosome positioning profile of *P* element insertion sequences from *Drosophila melanogaster*. The profile represents a model that analyzed 795 insertion sites (*P* flank 200) depicted by the black line and 2454 random DNA sequences (*P* random 200) depicted as the gray line. The X-axis represents nucleotide position while the Y-axis represents the unsigned nucleosome positioning values corresponding to absolute or only positive values. Flexibility is inferred from values closer to zero whereas rigidity is implied when values are more positive. A clear difference in trends is observed between the signals from the insertion sites and that of the random DNA. Comparable to the trends seen in the case of signed nucleosome positioning, this absolute scale also indicates that there are visible transitions in patterns going from less bendable or rigid regions of DNA upstream of the insertion point to more bendable regions very close to the insertion point and then back to rigid regions of DNA downstream of the insertion point. The random DNA sequences reveal less flexibility at the consensus sequence.

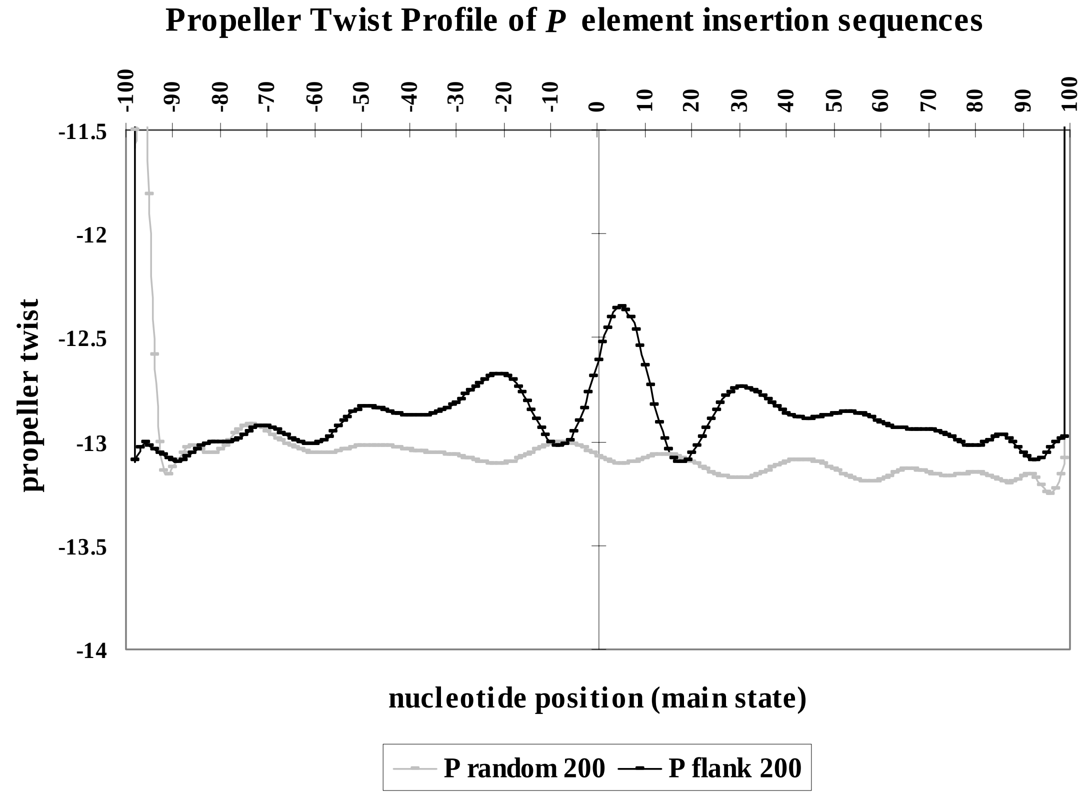


Figure VI-4: DNA propeller twist profile of *P* element insertion sequences from *Drosophila melanogaster*. 795 insertion sites are represented by the black line (*P* flank 200) 2454 random DNA sequences are depicted by the gray line (*P* random 200). The X-axis gives nucleotide position ranging from -100 to +100 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis is a measure of propeller twist. Flexibility is inferred from less negative values whereas rigidity is implied when values are more negative. The signal from the *P* element insertion sites reveals a region of flexibility at the target site flanked by less flexible regions on either side. Also, the opposite trend is seen with the random DNA and clearly at the start of the insertion consensus sequence, the two signals appear to exhibit trends in opposite directions.

The *P* element stacking energy profile is shown in figure V1-5. Stacking energy is given by a dinucleotide scale and is a measure of DNA stability. All values in this scale are negative. Less negative values correspond to regions in the helix that are able to melt or de-stack more readily and hence infer instability in those regions whereas more negative values correspond to regions in the helix that do not melt or de-stack readily due to a locking effect and hence correspond to highly stable regions. It is seen in figure V1-5 that the signal from the insertion sites has an average stacking energy of about -8.07592 at the target site whereas the signal from the random DNA has an average stacking energy of about -7.72708 at the target site. Hence it is clear that the region in the vicinity of the target site is more stable in the case of the insertion sites than that of the random DNA. The signal from the insertion sites transitions from unstable regions just before the start of the target site to a stable region at the target site and is then followed again by a region of instability.

A weblogo profile was generated to illustrate the information content of the *P* element insertion sites (Figure VI-6). Similar to the previous profiles, the start of the target site is denoted as position 0 and upstream and downstream sequences are denoted as '-' and '+' respectively. Figure VI-6 shows a fairly consistent 8bp GGCCAGAC consensus pattern as reported previously. The sequence logo of the *P* element also suggests that a strong preference for A/T nucleotides seems likely at exactly 3 bps (-3 and +10) from the start (position 0) and end (position +7) of the 8bp consensus sequence. Flanking sequences upstream and downstream of positions -3 and +10 show no preference for any nucleotide. It is possible that the information content of the 3 bps immediately preceding and following the consensus sequence at the insertion site contributes to the selection of these sites for element integration.

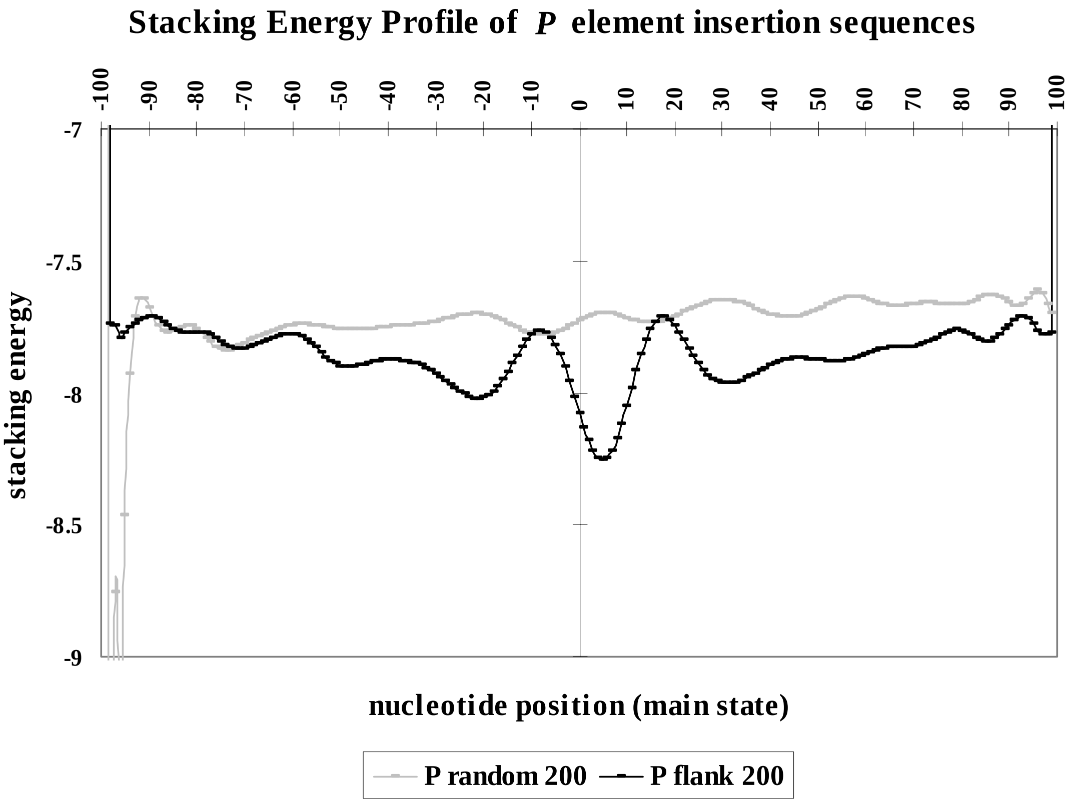


Figure VI-5: Stacking energy profile of *P* element insertion sequences from *Drosophila melanogaster*. 795 insertion sites are represented by the black line (*P* flank 200) 2454 random DNA sequences are depicted by the gray line (*P* random 200). The X-axis gives nucleotide position ranging from -100 to +100 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis is a measure of stacking energy. Stability is inferred from less negative values whereas instability is implied when values are more negative. It is seen in this figure that the *P* element insertion sites reveal more stability at the target site than the random DNA.

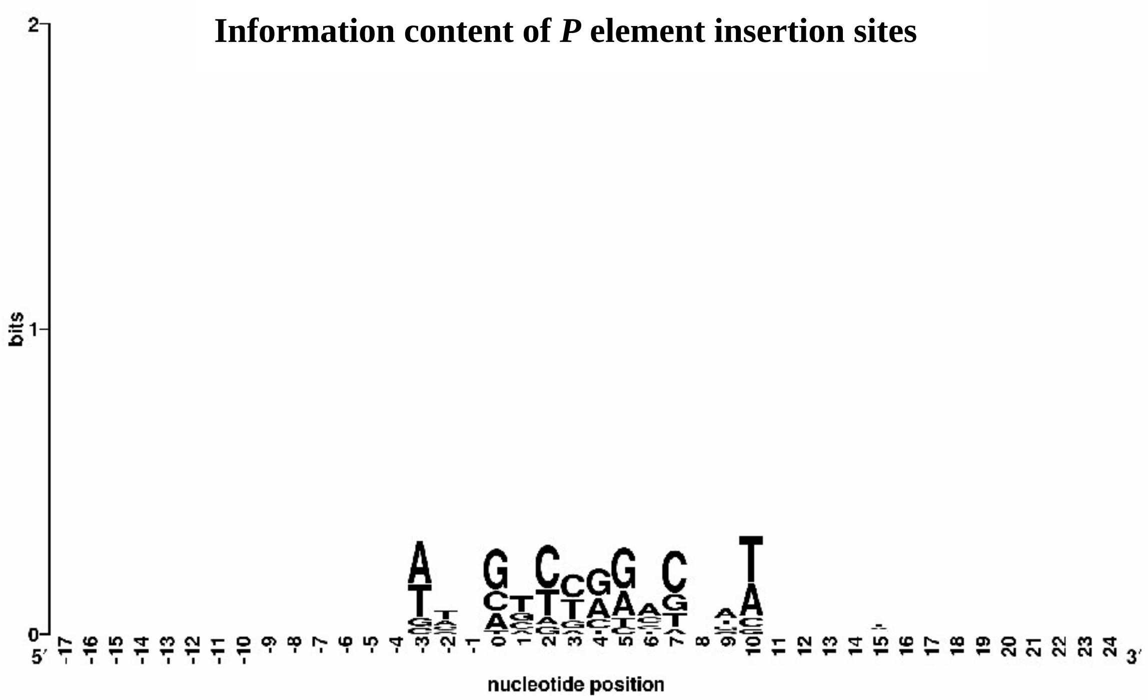


Figure VI-6: Sequence logo of 795 *P* element insertion sites generated by the Weblogo program. The X-axis depicts the nucleotide position of the insertion site sequences and ranges from -17 to +24 wherein the insertion site is at base 0. The Y-axis is given in bits with 2 being the maximum number of bits. There could be a combination of alphabets or a dominance of a particular alphabet at any nucleotide position and hence the height of each alphabet is a measure of its relative occurrence at that position. The 8bp consensus sequence (0 to +7) has 17 bps of flanking sequences upstream (+8 to +24) and downstream (-1 to -17) of the insertion site. There is a preference for A/T nucleotides at the 3rd bp before (position -3) and after (position +10) the start and end of the target consensus sequence, respectively.

6.2 Profiles of *Tc1* element insertions in the *Caenorhabditis elegans* genome

Contrary to the trends seen in the *P* element profiles, the first set of *Tc1* element insertion site sequences do not show distinguishable patterns of bendability or flexibility at the target site consensus sequence. The bendability profile of the *Tc1* element as shown in Figure VI-7 portrays a very random pattern of peaks all throughout the 200 bps of flanking sequence.

As opposed to the *P* element, which revealed a clear change in trend with respect to DNA being more bendable at and very near to the insertion point compared to the surrounding flanking DNA, the *Tc1* element does not show pronounced peaks or changes in trends in the vicinity of the insertion point alone, but instead reveals a fairly random bendability profile. The average value of bendability at the insertion point was -0.0312563.

The second set of *Tc1* insertions within the gpa-2 gene that were analyzed was similar to the first dataset. The bendability profile of this dataset also revealed a very random pattern of trends and no clearly distinguishable pattern of bendability could be identified at the insertion site (Figure VI-8). The average value of bendability for this dataset was -0.0326517, which was consistent with that of the first dataset.

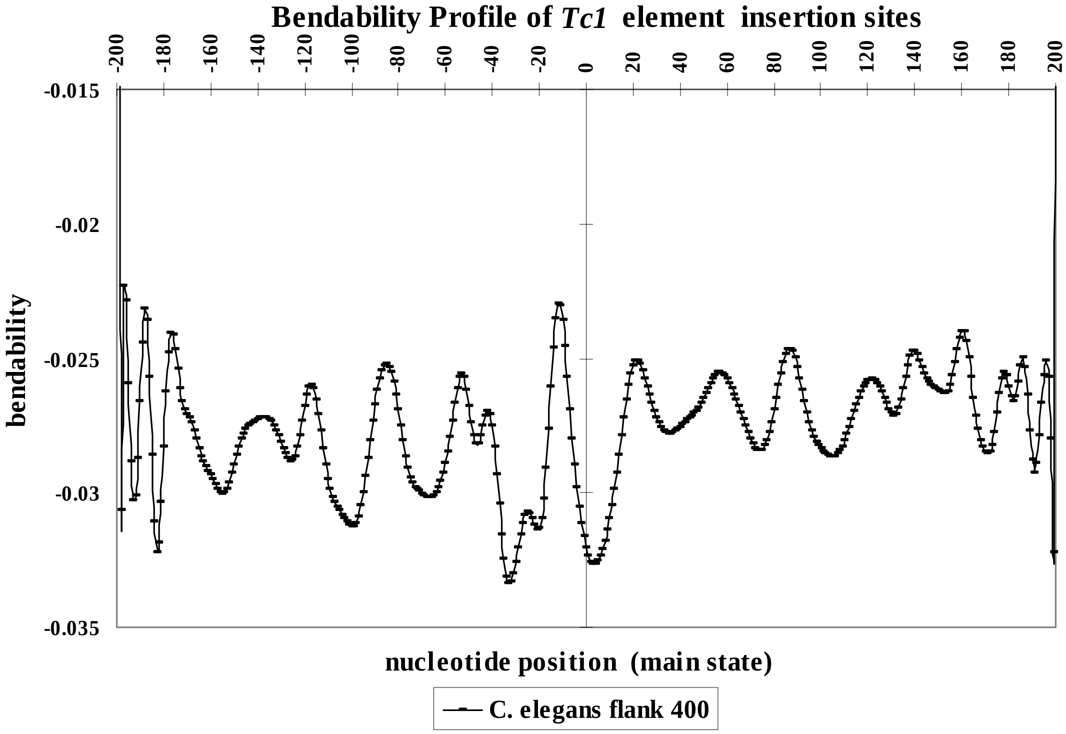


Figure VI-7: DNA bendability profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The profile represents a model that analyzed 196 insertion sites (*C.elegans* flank 400) depicted by the black line. The X-axis represents nucleotide position ranging from -200 to +200 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the bendability values. Bendability is a measure of anisotropic flexibility and is given by a trinucleotide scale. The more positive or closer the value to zero implies a greater degree of bendability, whereas more negative values indicate rigidity. As seen in this figure, the signal from the *Tc1* insertion sites is quite random and no distinct trends can be observed in and around the target site as opposed to the trends seen in the structural profiles of the *P* element.

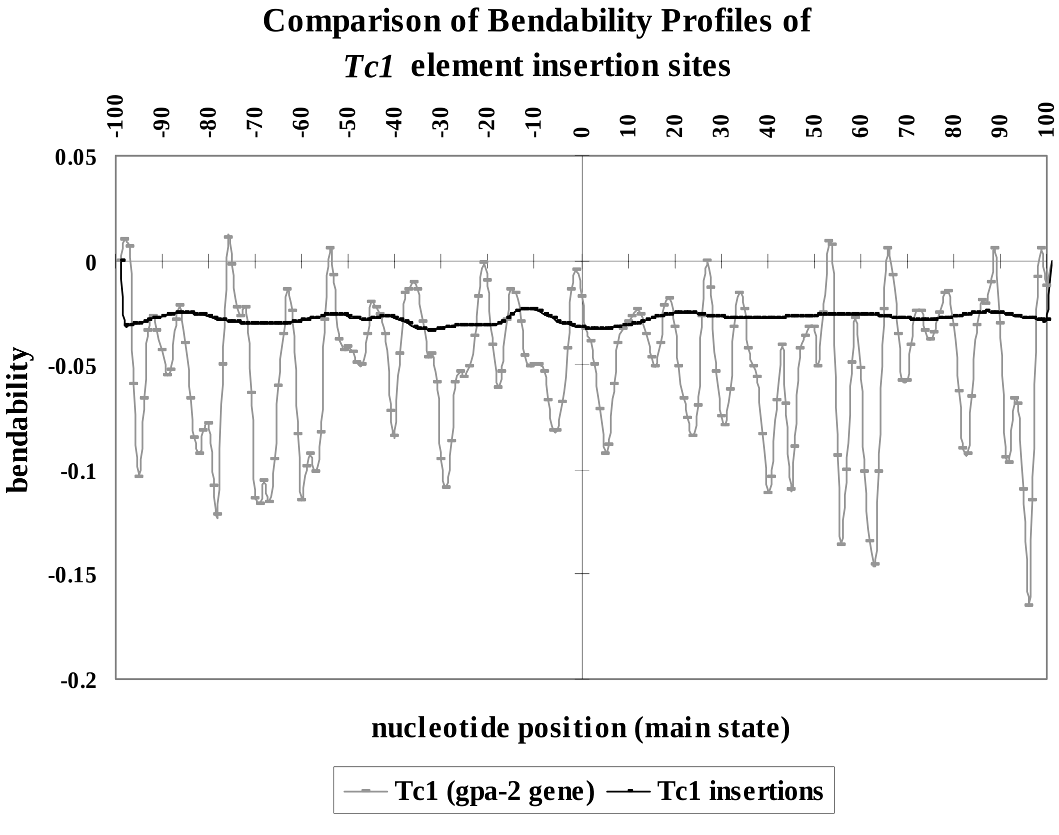


Figure VI-8: Comparison of DNA bendability profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The first set of 196 *Tc1* element insertions is shown in black and the second set of 22 *Tc1* insertions identified within the gpa-2 gene is shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the trinucleotide bendability values. The more positive or closer the value to zero implies a greater degree of bendability, whereas more negative values indicate rigidity. As seen in this figure, the signals from both sets of *Tc1* insertion sites is quite random and no distinct trends can be observed in and around the target site as opposed to the trends seen in the bendability profile of the *P* element.

Figure VI-9 reveals the signed nucleosome positioning profile of the first dataset of *Tc1* element insertions while figure VI-10 gives a comparison between the first and second sets of *Tc1* element insertions. Both figures reveal patterns of varying degrees of flexibility and rigidity across the entire profile. The *Tc1* element actually indicates the lowest nucleosome positioning value at the target site, although the high fluctuation in signal throughout the profile renders this profile insignificant. Unlike the *P* element that showed a distinct change in degree of flexibility at the insertion site, the *Tc1* element profile did not reveal any distinct trends in flexibility either at or immediately around the insertion point.

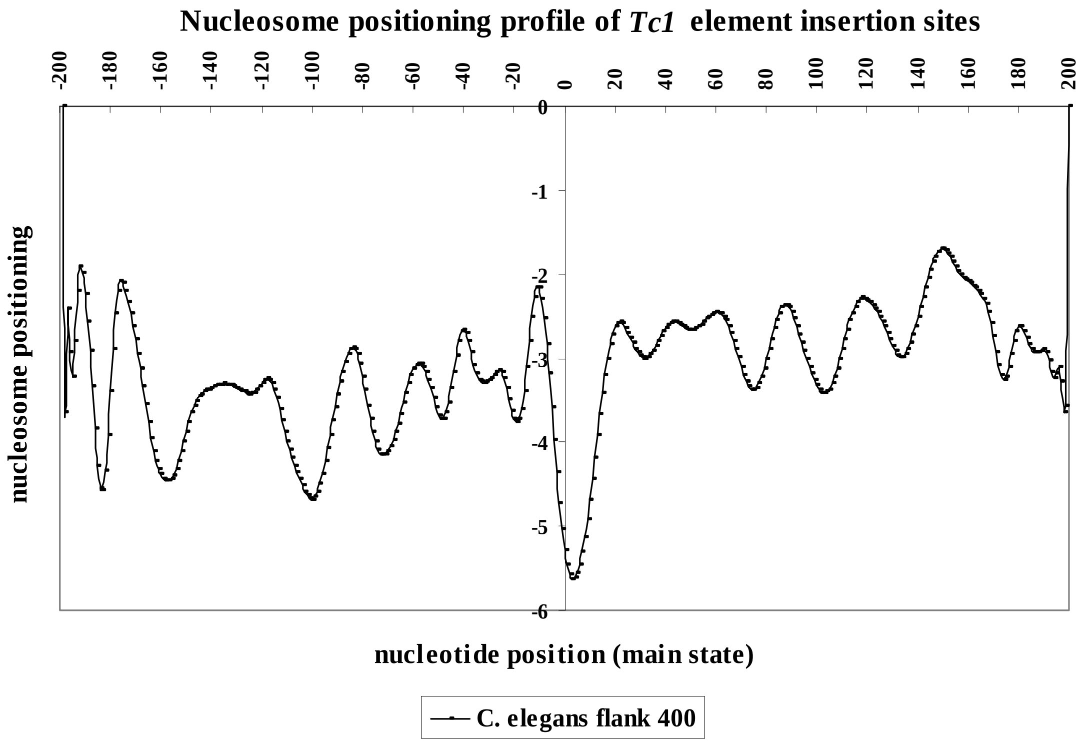


Figure VI-9: DNA nucleosome positioning profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. Profile represents a model that analyzed 196 insertion sites (*C.elegans* flank 400) depicted by the black line. The X-axis represents nucleotide position ranging from -200 to +200 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis represents the signed nucleosome positioning values. The higher or more positive the value of nucleosome positioning, the greater the flexibility of DNA in that region. The trend of the *Tc1* insertions indicates regions of variable flexibility and rigidity. No significant patterns of flexibility are seen at the insertion site with respect to surrounding regions.

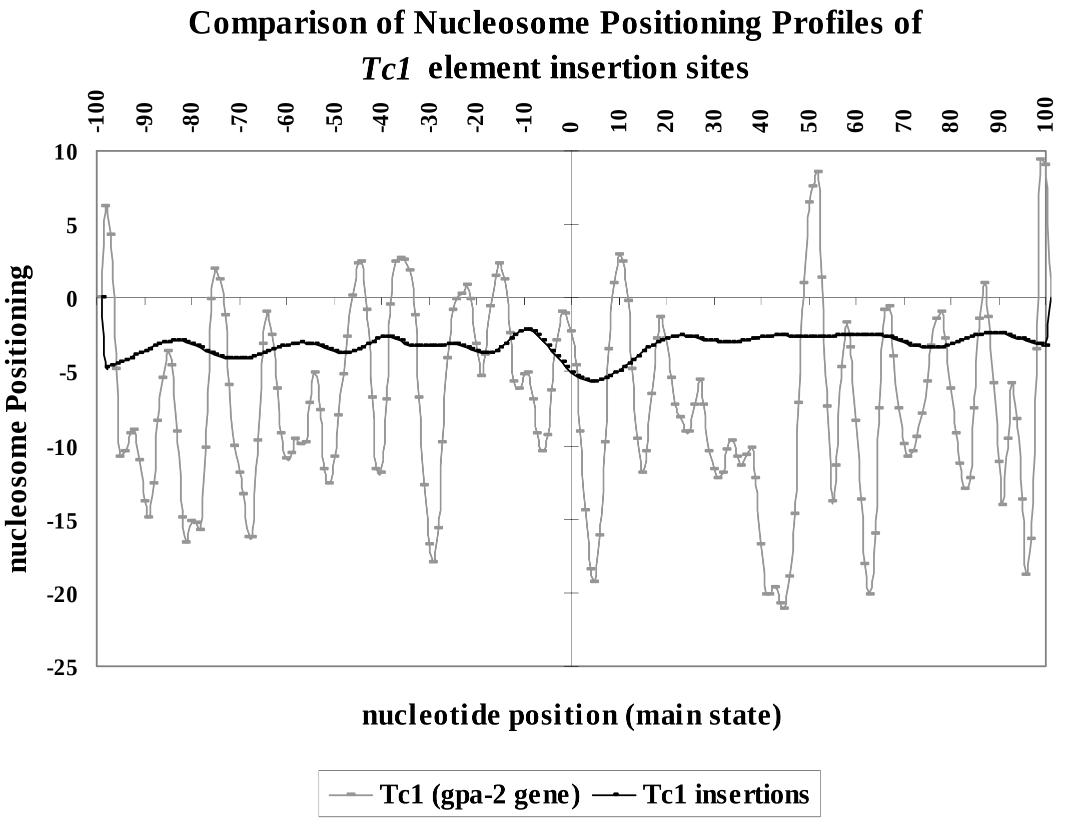


Figure VI-10: Comparison of DNA nucleosome positioning profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The first set of 196 *Tc1* element insertions is shown in black and the second set of 22 *Tc1* insertions identified within the gpa-2 gene is shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the trinucleotide nucleosome positioning values. This is an absolute scale and the closer the value to zero, the more flexible, whereas more positive values indicate rigidity. As seen in this figure, the signals from both sets of *Tc1* insertion sites is quite random and no distinct trends can be observed in and around the target site as opposed to the trends seen in the bendability profile of the *P* element.

Figure VI-11 reveals the unsigned nucleosome positioning profile of the first dataset of *Tc1* element insertions while figure VI-12 gives a comparison between the first and second sets of *Tc1* element insertions. These profiles are comparable to the signed nucleosome positioning profiles in that they exhibit extremely varying patterns upstream and downstream of the target site. At the target site, it is seen that the *Tc1* element insertion sites has the highest positive value of 15.32204 and suggests rigidity. However, this does not seem significant in comparison to the wide variation in signal all through the profile.

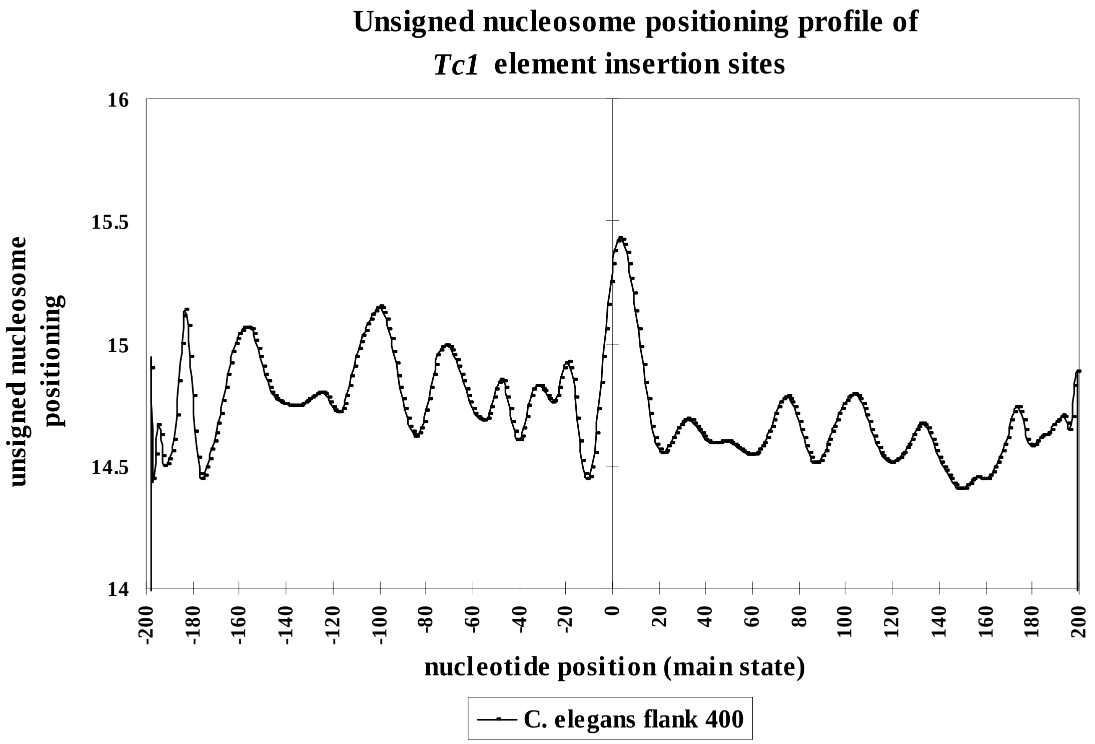


Figure VI-11: DNA unsigned nucleosome positioning profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The profile represents a model that analyzed 196 insertion sites (*C.elegans* flank 400) depicted by the black line. The X-axis represents nucleotide position ranging from -200 to +200 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis represents the signed nucleosome positioning values. The higher or more positive the value of nucleosome positioning, the greater the rigidity of DNA in that region. The trend of the *Tc1* insertions indicates regions of variable flexibility and rigidity. No significant patterns of flexibility are seen at the insertion site with respect to surrounding regions.

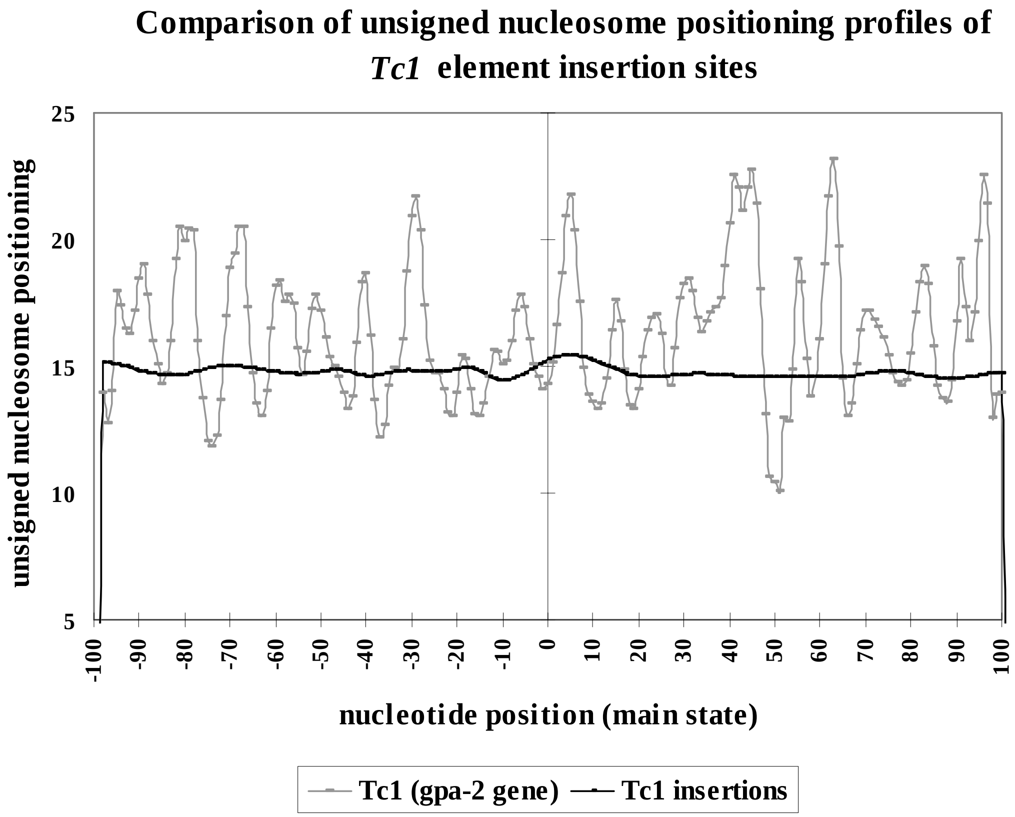


Figure VI-12: Comparison of DNA unsigned nucleosome positioning profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The first set of 196 *Tc1* element insertions is shown in black and the second set of 22 *Tc1* insertions identified within the gpa-2 gene is shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the trinucleotide nucleosome positioning values. This is an absolute scale and the closer the value to zero, the more flexible, whereas more positive values indicate rigidity. As seen in this figure, the signals from both sets of *Tc1* insertion sites is quite random and no distinct trends can be observed in and around the target site as opposed to the trends seen in the bendability profile of the *P* element.

Figure VI-13 reveals the propeller twist profile of the first dataset of *Tc1* element insertions while figure VI-14 gives a comparison between the first and second sets of *Tc1* element insertions. These profiles are comparable to the previously described bendability and nucleosome positioning profiles of the *Tc1* element. As seen previously, the profiles are random and fast changing trends varying between regions of flexibility and rigidity. However, in keeping with the other structural parameters examined, the target site at base 0 exhibits one of the highest values of -13.8453 indicating rigidity.

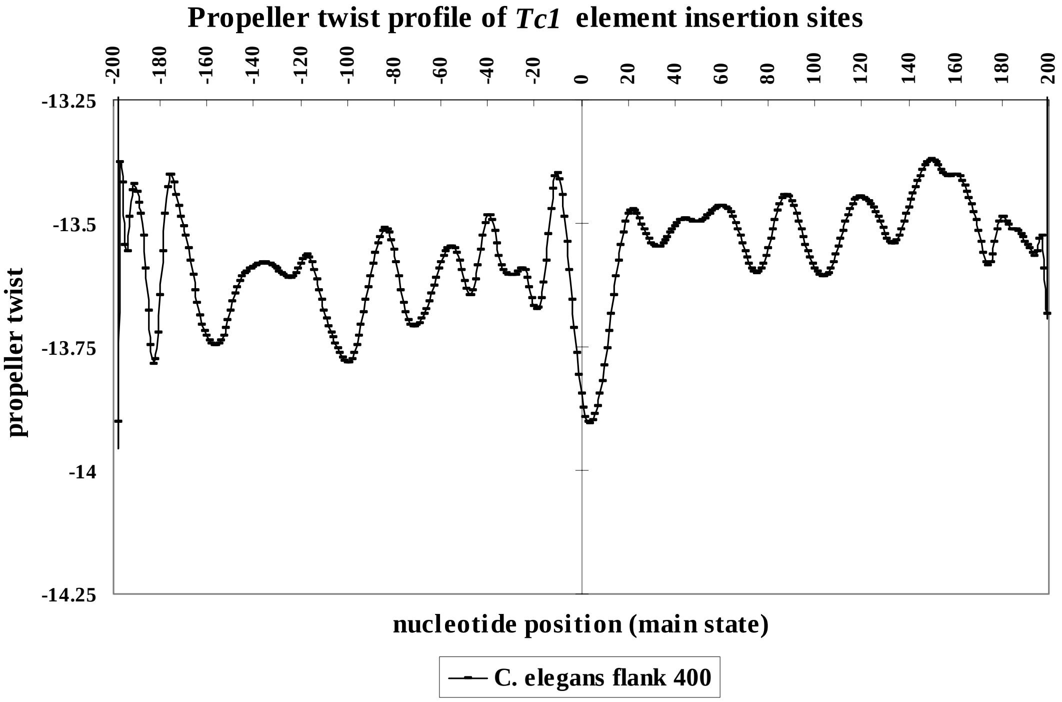


Figure VI-13: DNA propeller twist profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. This figure represents the first set of 196 *Tc1* element insertions shown in black. The X-axis gives nucleotide position ranging from -200 to +200 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis is a measure of propeller twist. Flexibility is inferred from less negative values whereas rigidity is implied when values are more negative. The signal from the *Tc1* element insertion sites reveals rapid changes in trends from being bendable to rigid and vice versa. However, the signal at the target site indicates the highest propeller twist value and corresponds to rigidity.

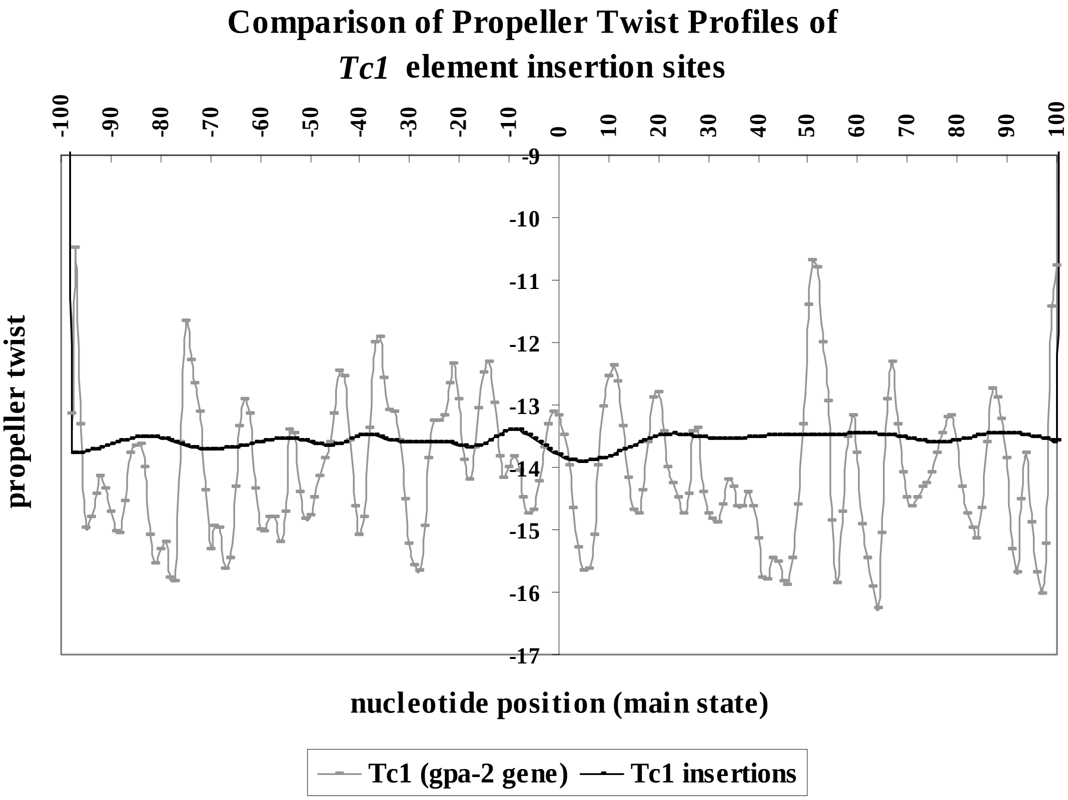


Figure VI-14 : Comparison of propeller twist profiles of Tc1 element insertion sites in the *Caenorhabditis elegans* genome. The first set of 196 Tc1 element insertions is shown in black and the second set of 22 Tc1 insertions identified within the gpa-2 gene is shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the dinucleotide propeller twist values. The signals from both sets of Tc1 insertion sites exhibit rapid variations in trends. However, the variations in the second dataset shown in gray appear more significant than the signal shown in black.

The *Tc1*element stacking energy profile is seen in figure VI-15. It indicates that the target site reveals a region of low stacking energy values (-7.13067) corresponding to instability and this is immediately flanked upstream and downstream by regions of higher stability. However, as seen in all the other *Tc1* profiles, this profile exhibits a lot of variation in trends throughout the profile. Figure VI-16 gives a comparison of stacking energy profiles of the two sets of *Tc1*elements. Here again it is seen that both signals are busy throughout the profile and do not serve to conclude much about the stability at the target site.

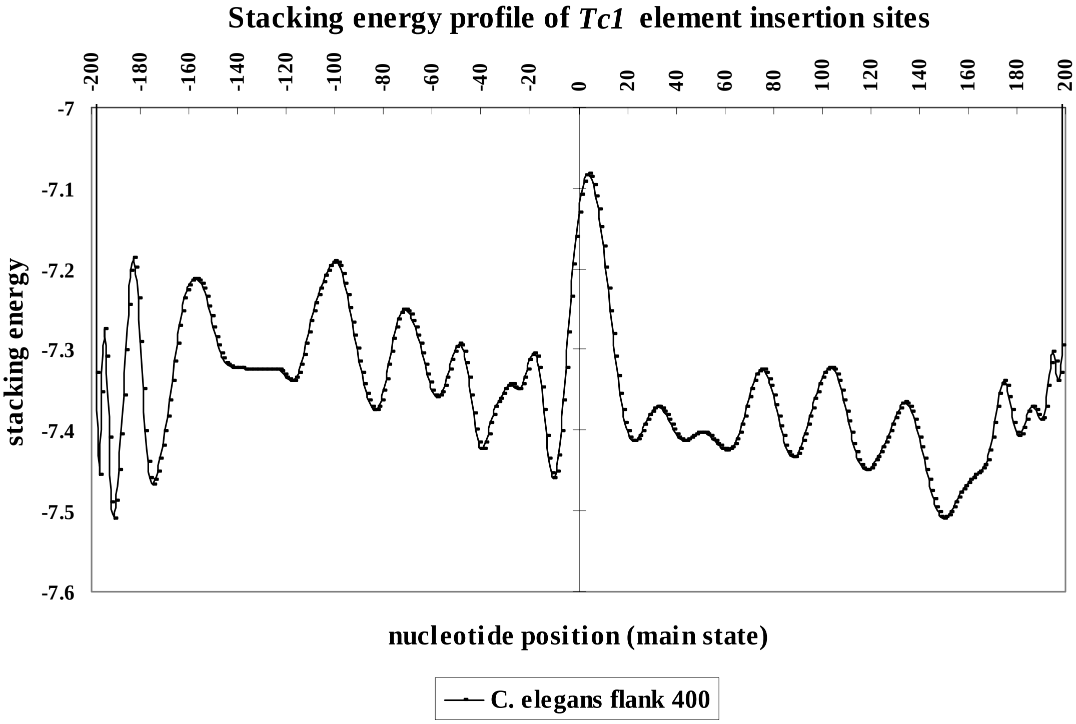


Figure VI-15: DNA stacking energy profile of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. This figure represents the first set of 196 *Tc1* element insertions shown in black. The X-axis gives nucleotide position ranging from -200 to +200 bps with the start of the consensus sequence at position 0 and flanking DNA on either side. The Y-axis is a measure of stacking energy. Stability is inferred from more negative values whereas rigidity is implied when values are less negative. The signal from the *Tc1* element insertion sites reveals rapid changes in trends from being bendable to rigid and vice versa. However, the lowest value for stacking energy is seen close to the target site and this indicates regional instability.

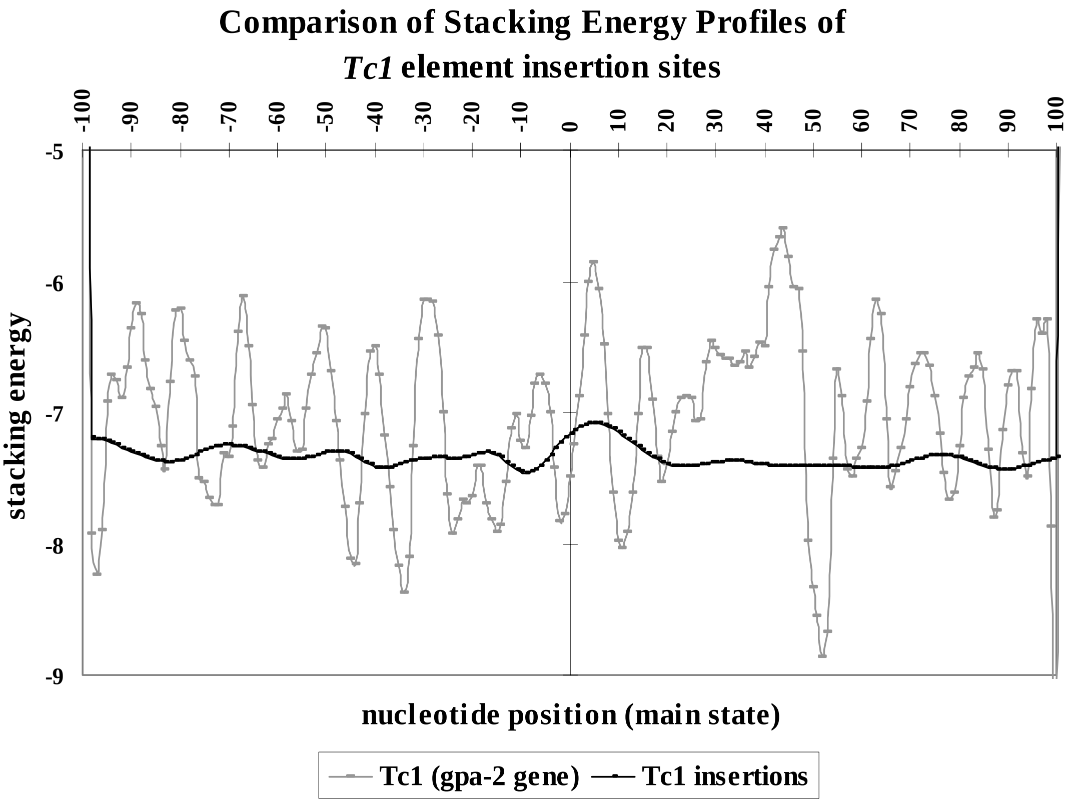


Figure VI-16: Comparison of stacking energy profiles of *Tc1* element insertion sites in the *Caenorhabditis elegans* genome. The first set of 196 *Tc1* element insertions is shown in black and the second set of 22 *Tc1* insertions identified within the gpa-2 gene is shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the 8 bp consensus target sequence at position 0 and flanking DNA on either side. The Y-axis represents the trinucleotide stacking energy values. The signals from the second set of *Tc1* insertion sites shown in gray exhibit rapid variations in trends, however, the signal from the first set shown in black seems quite symmetrical about the insertion site. Also, for the first set, the lowest values of stacking energy are seen in and around the centre of the target site and this depicts instability in that region.

A weblogo profile was generated to illustrate the information content of the first (Figure VI-17) and second sets of *Tc1* element insertion sites (Figure VI-18). Similar to the structural profiles, the start of the target site is denoted as position 0 and upstream and downstream sequences are denoted as ‘-’ and ‘+’ respectively. Both figures represent the target site consensus TA dinucleotide at positions 0 and 1 along with information content at 10 bps upstream and downstream of the target site.

The sequence logo of the first set (Figure VI-17) suggests that a strong preference for A/T nucleotides seems likely at exactly 3 bps (-3 and +4) from the start (position 0) and end (position +1) of the TA consensus sequence. Flanking sequences upstream and downstream of positions -3 and +4 show no preference for any nucleotide. It was also observed that there is no preference for any nucleotide at exactly 4 bps upstream (position -4) and downstream (position +5) of the target consensus sequence.

However, the weblogo profile of the second dataset (Figure VI-18) revealed a different pattern of occurrence of nucleotides in and around the target site. Since the second dataset only represented 22 insertions as opposed to the first dataset consisting of 196 insertions, it is possible that the second dataset is not a good representation of the true consensus patterns for the *Tc1* element insertions.

Therefore, based on the weblogo profile of the first dataset, it is possible that the information content of the 3 bps immediately preceding and following the consensus sequence at the insertion site and also the lack of preference for any nucleotide at 4 bps up/downstream of the target site contributes to the selection of these sites for element integration.

Information content of *Tc1* element insertion sites

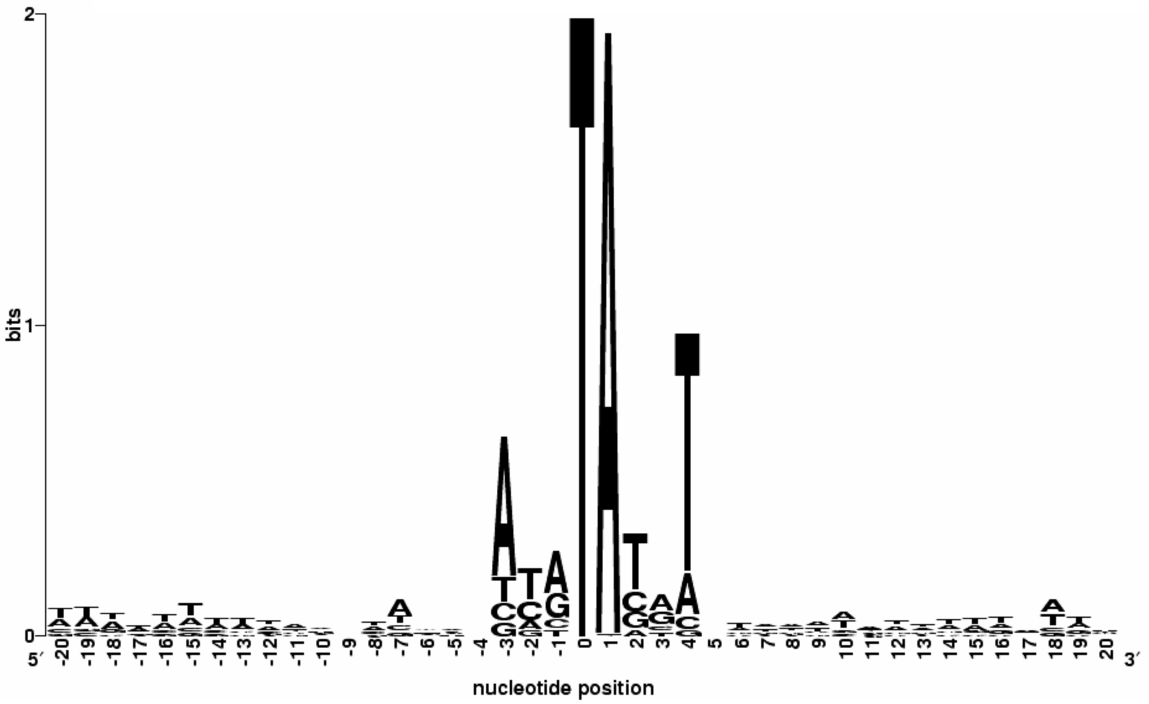


Figure VI-17: Sequence logo of 196 *Tc1* element insertion sites generated by the Weblogo program. The X-axis depicts the nucleotide position of the insertion site sequences and ranges from -17 to +24 wherein the insertion site is at base 0. The Y-axis is given in bits with 2 being the maximum number of bits. There could be a combination of alphabets or a dominance of a particular alphabet at any nucleotide position and hence the height of each alphabet is a measure of its relative occurrence at that position. The target consensus sequence (0 to +1) corresponds to TA dinucleotides and this is flanked by 18 bps of upstream sequences (+2 to +20) and 20 bps of downstream sequences. There is a preference for A/T nucleotides at the 3rd bp before (position -3) and after (position +4) the start and end of the target consensus sequence, respectively. There is no preference for any nucleotide at the 4th bp before (position -4) and after (position +5) the start and end of the target consensus sequence.

Information content of *Tc1* element insertion sites within the gpa-2-gene

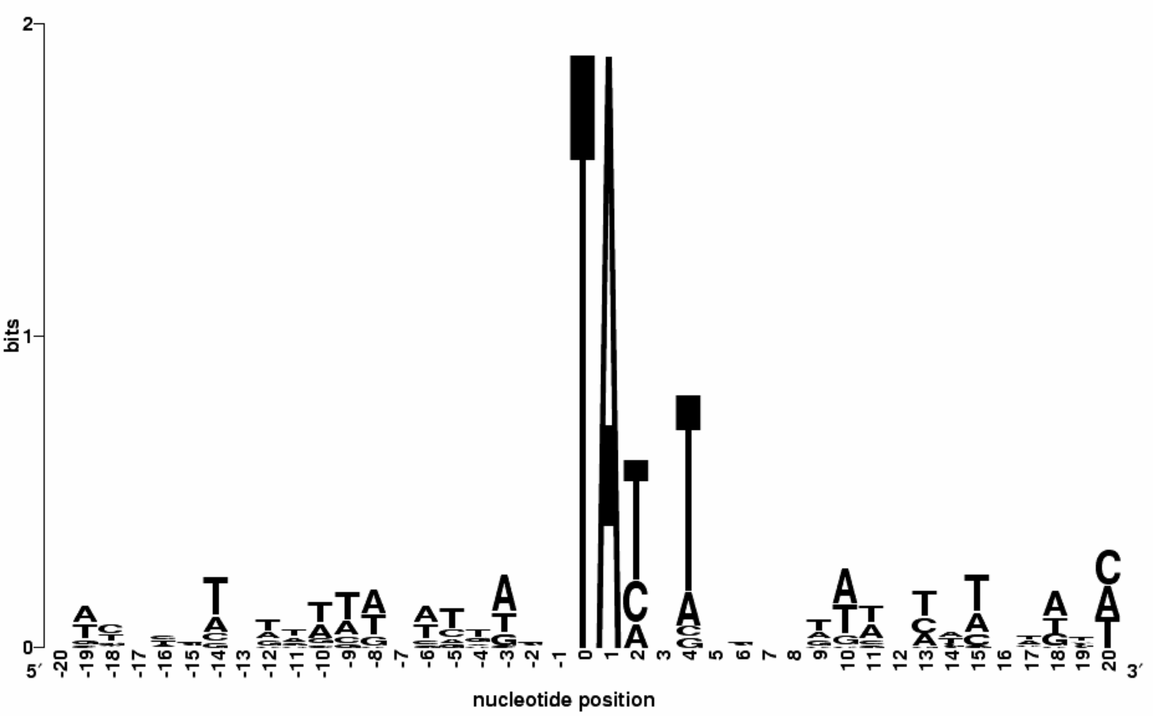


Figure VI-18: Sequence logo of the second set of 22 Tc1 element insertion sites generated by the Weblogo program. The X-axis depicts the nucleotide position of the insertion site sequences and ranges from -17 to +24 wherein the insertion site is at base 0. The Y-axis is given in bits with 2 being the maximum number of bits. There could be a combination of alphabets or a dominance of a particular alphabet at any nucleotide position and hence the height of each alphabet is a measure of its relative occurrence at that position. The target consensus sequence (0 to +1) corresponds to TA dinucleotides and this is flanked by 18 bps of upstream sequences (+2 to +20) and 20 bps of downstream sequences (-1 to -20). There is a preference for A/T nucleotides at the 1st bp (position +2) and 3rd bp (position +4) after the end of the target consensus sequence, whereas there is no preference for any nucleotide at the 2nd bp (position +3) after the end of the target sequence. Also, there is no preference for any nucleotide at the 1st bp (position -1) before the start of the TA dinucleotide.

6.3 Profiles of *Hermes* element insertion sites in the pGDV1 target plasmid

Similar to the analysis performed with the *P* and the *Tc1* elements, the *Hermes* element was also investigated for any similarities or differences in structural properties between insertion sites and consensus sequences that were not hit (cold spots). This was the first element investigated using plasmid based transposition assays.

As in the case of the *P* element, the DNA bendability profile for *Hermes* insertion sites yielded profiles that suggested definitive changes in trends. Insertion sites tended to be bendable regions, whereas the cold spots showed a lesser degree of bendability at the consensus sequence. The DNA bendability profile is shown in Figure VI-19.

The middle of the consensus target sites (position +4) shows an average value of -0.0360483 while the cold spots (position +4) have an average value of -0.0254135. While this difference was not drastic, the sequences located immediately upstream of the insertion site showed the largest difference in bendability between insertion sites and cold spots, perhaps explaining the orientation preference of the *Hermes* element into hot spots.

Figure VI-20 illustrates the second structural parameter that was taken into consideration in this analysis, namely nucleosome positioning, with this profile also being consistent with insertion sites showing more flexibility than the cold spots. Furthermore, it is even more distinguishable in this profile due to the wide difference in values with the insertion sites (-2.95143) and the cold spots (-11.3197).

Figure VI-21 illustrates the unsigned nucleosome positioning profile which uses an absolute scale. Similar to the signed nucleosome positioning profile, it is clear that there is a remarkable difference in values at the target site between the insertion sites (14.42141) being more flexible and the cold spots (17.75642) being less flexible.

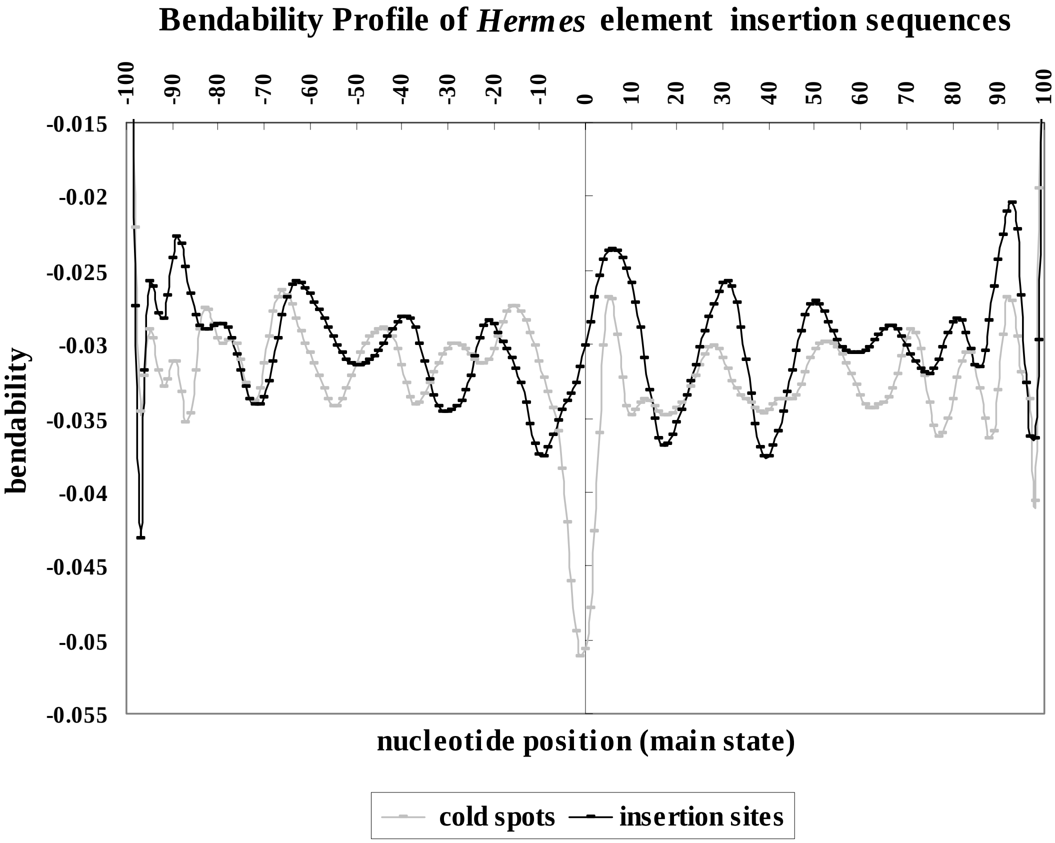


Figure VI-19: DNA bendability profile for *Hermes* element insertion sequences in the pGDV1 target plasmid. 76 insertion sites were analyzed as shown by the black line and 51 cold spots shown in gray. The X-axis depicts the nucleotide position with the 8 bp consensus target sequence starting at base 0 and flanking sequences on either side. Y-axis gives the bendability scales which are based on trinucleotide values ranging from -0.280 (rigid) to +0.194 (very bendable). There is an easily observable trend in variability of bendability around the consensus target site in both cases, but the most difference in bendability values between the insertion sites and the cold spots is seen 5 to 8 bps upstream of the target consensus sequence. For the insertion sites, there is a gradual trend of DNA being more bendable between positions -10 and 0. Simultaneously, within approximately the same region between positions -10 and 0 it is observed that for the cold spots, there is an initial downward trend of DNA being less bendable.

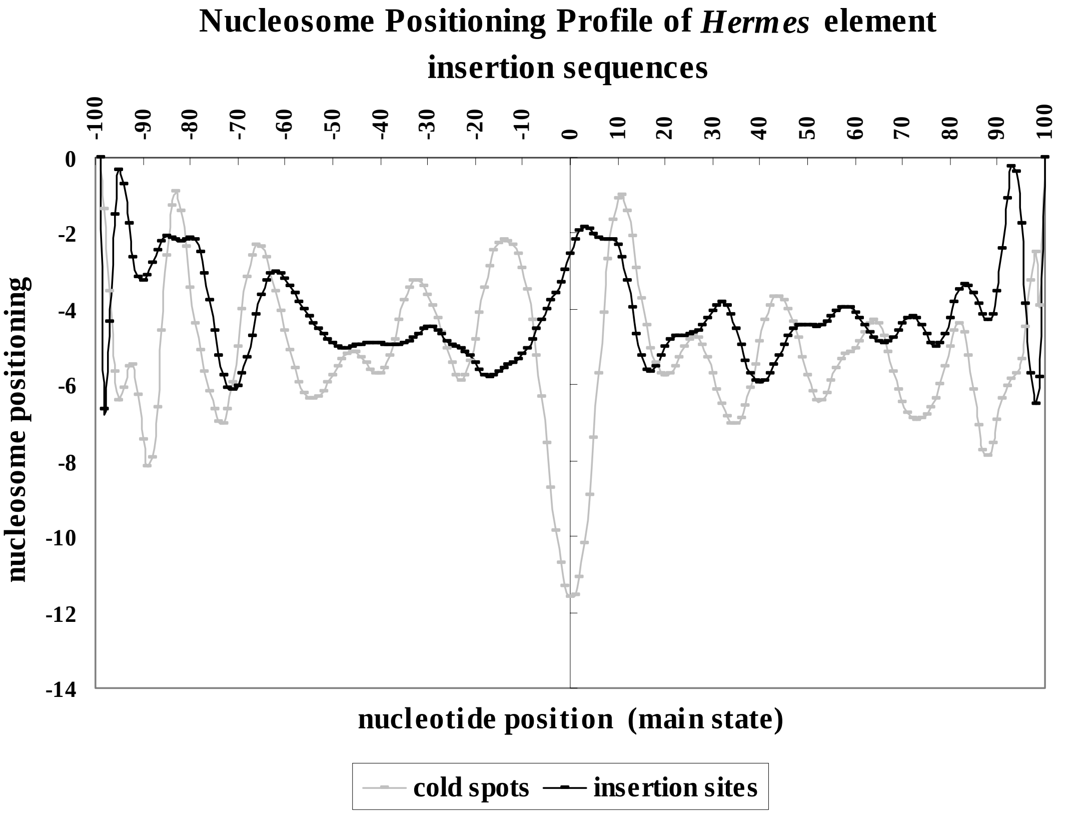


Figure VI-20: Nucleosome positioning profile of *Hermes* element insertion sequences. Here again 76 insertion sites and 51 cold spots were analyzed. X-axis depicts the nucleotide position with the consensus target sequence starting at base 0 and flanking sequences are on either side. Y-axis gives the nucleosome positioning scales ranging from -36 (straight, rigid) to +45 (strong bends). There is a variation in signal going from less bendable approximately 15 bps before the insertion point to most flexible close to the insertion consensus sequence and then back to less bendable about 10 bps after the insertion point. Cold spots show the lowest value of approximately –12 at about the insertion point, suggesting much less DNA flexibility than the insertion sites. The cold spots exhibit trends contrary to the signal at the insertion sites, going from most bendable approximately 15 bps before the consensus sequence to least bendable near the target consensus sequence and finally back to most bendable after the consensus sequence.

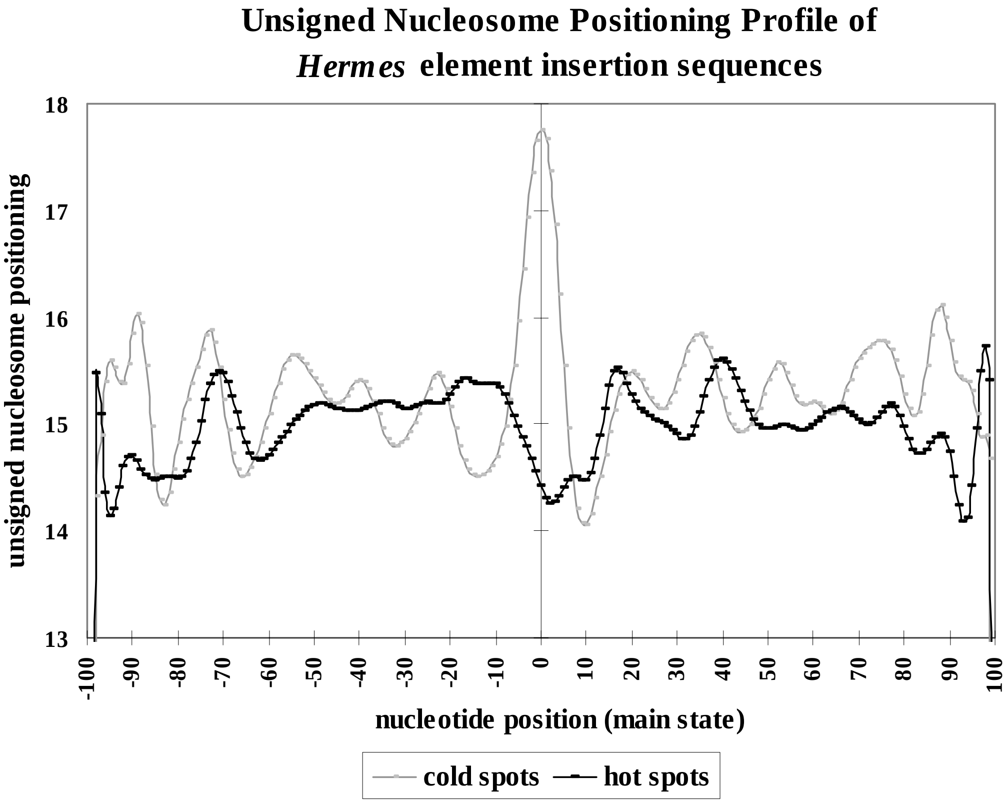


Figure VI-21: Unsigned nucleosome positioning profile of *Hermes* element insertion sequences. 76 insertion sites and 51 cold spots were analyzed in this model. X-axis depicts the nucleotide position with the consensus target sequence starting at base 0 and flanking sequences are on either side. Y-axis gives the unsigned nucleosome positioning scales. There is a variation in signal going from less bendable approximately 15 bps before the insertion point to most flexible close to the insertion consensus sequence and then back to less bendable about 10 bps after the insertion point. Cold spots show the highest positive value of approximately 17.8 at about the insertion point, suggesting much less DNA flexibility than the insertion sites. The cold spots exhibit trends contrary to the signal at the insertion sites, going from most bendable approximately 15 bps before the consensus sequence to least bendable near the target consensus sequence and finally back to most bendable after the consensus sequence.

Figure VI-22 illustrates propeller twist of the *Hermes* element insertion sequences. This profile is a good illustration of the differences in flexibility between the insertion sites and the cold spots. It is clearly seen that the largest difference in trends is at the target site where the insertion sites reveal a low propeller twist value of -13.4547 indicating higher flexibility, whereas, the cold spots revealed a high twist of -14.7002 corresponding to lower flexibility.

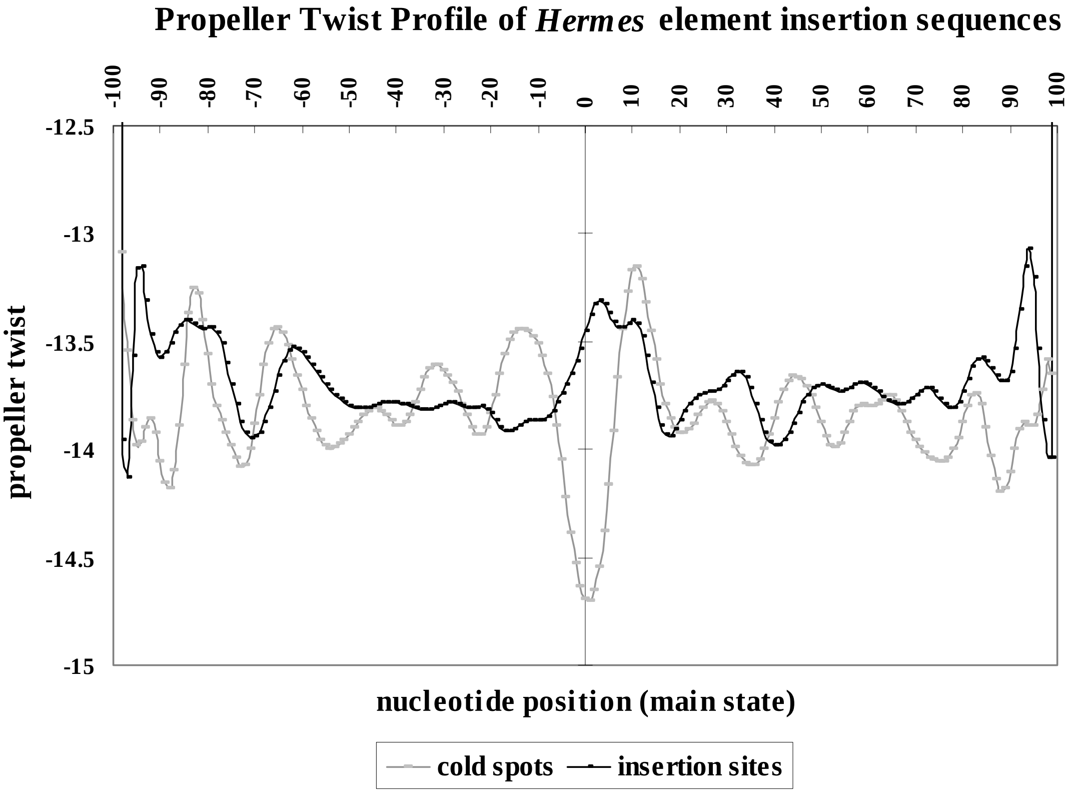


Figure VI-22: Propeller twist profile of *Hermes* element insertion sequences. 76 insertion sites and 51 cold spots were analyzed in this model. X-axis depicts the nucleotide position with the consensus target sequence starting at base 0 and flanking sequences are on either side. Y-axis gives a measure of propeller twist. There is a variation in signal going from less bendable approximately 10 bps before the insertion point to most flexible close to the insertion consensus sequence and then back to less bendable about 18 bps after the insertion point. The cold spots exhibit trends contrary to the signal at the insertion sites, going from most bendable approximately 10 bps before the consensus sequence to least bendable near the target consensus sequence and finally back to most bendable after the consensus sequence. The greatest difference in signal trends is seen at the target site where a definite opposition of trends is visible.

Figure VI-23 illustrates stacking energy of the *Hermes* element insertion ssequences. This profile is indicates differences in stability between the two signals. It is clearly again seen that the largest difference in trends is at the target site where the insertion sites reveal a high stacking energy (high stability) of -7.45707, whereas, the cold spots revealed a low stacking energy (low stability) of -6.54005.

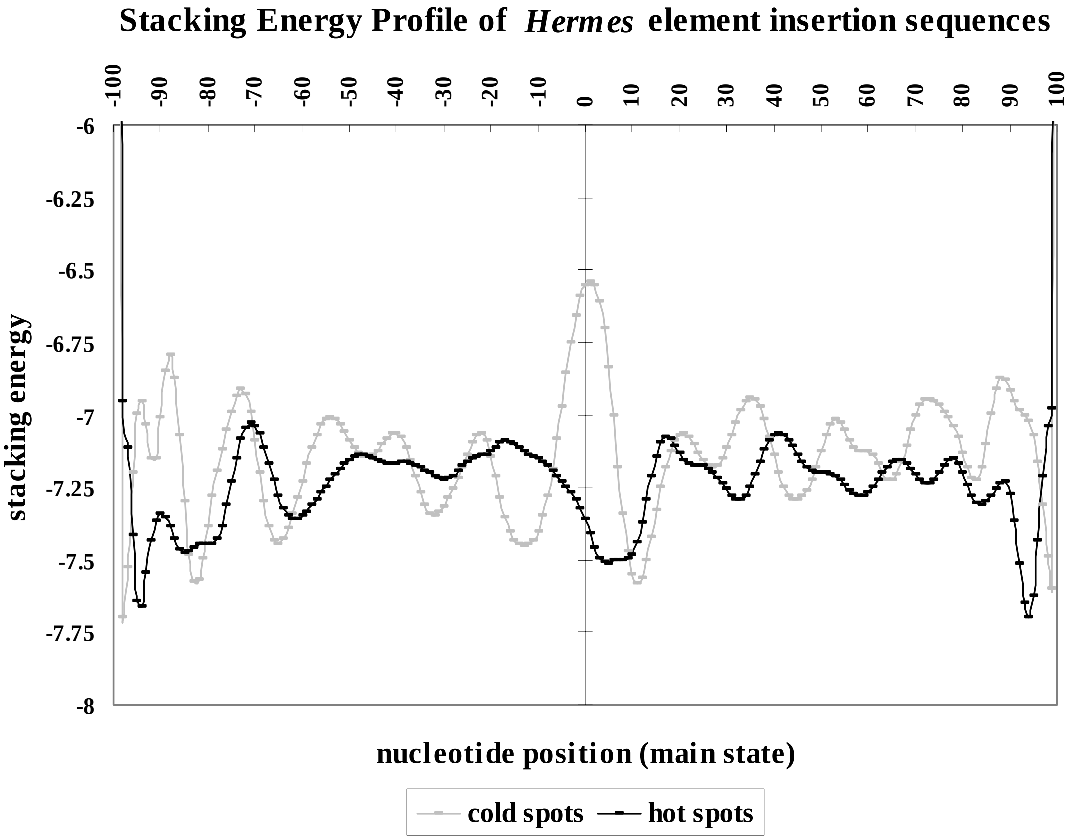


Figure VI-23: Stacking energy profile of *Hermes* element insertion sequences. 76 insertion sites and 51 cold spots were analyzed in this model. X-axis depicts the nucleotide position with the consensus target sequence starting at base 0 and flanking sequences are on either side. Y-axis gives a measure of stacking energy. The greatest difference in signal trends is seen at the target site where a definite opposition of trends is visible. The cold spots have the lowest stacking energy and the insertion sites have the highest stacking energy at the target site. This indicates that the cold spots are in a region of lower stability than the insertion sites.

The sequence logos of the insertion sites (Figure VI-24) and the cold spots (Figure VI-25) of the *Hermes* element are given. The sequence logo giving the information content of the insertion sites shows the 8 bp NTNNNNAC consensus sequence to a great extent, although the 8th bp of this consensus sequence is not well conserved. Based on the 76 insertion site sequences that were analyzed, there was 72% conservation of the 2nd bp, 74% conservation of the 7th bp and 36% conservation of the 8th bp. This occurs because the insertion sites for the *Hermes* element match 4 different yet related consensus patterns.

The 8 bp NTNNNNAC consensus sequence is better identified in the cold spots as these were manually identified using that particular consensus sequence. Of most interest are the nucleotides immediately upstream of the target sequence. The greatest information content for insertion sites occurs at position -2 with a preference for A/T nucleotides, whereas for the cold spots there is very little information content at position -2 and an increased content at positions -1 and -3 with a preference for T/A nucleotides. It is possible that the combination of these differences accounts for the insertion preference into the observed insertion sites rather than the unused cold spots.

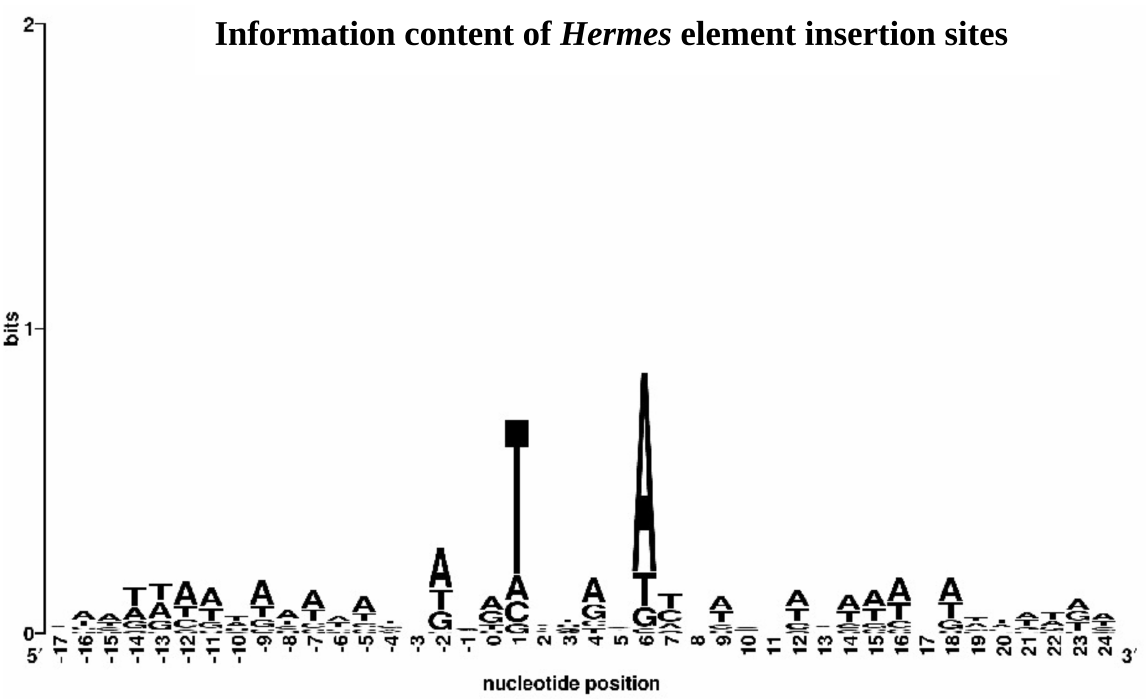


Figure VI-24: Sequence logo providing the information content of the insertion sites in pGDV1 for the *Hermes* element. The 8bp NTNNNNAC consensus target sequence starts at base 0 and as observed in this sequence logo, the consensus pattern is conserved. It also gives the information content of 17 bps of flanking sequence on either side (-17 to -1 and +8 to +24) of the insertion site. It is clear that there is no preference for any nucleotide at position -1 preceding the start of the consensus sequence and at position +8 immediately following the end of the consensus sequence. There also seems to be likelihood for A/T nucleotides at positions -2 and +9 and then general AT richness is seen in the DNA upstream and downstream of positions -2 and +9 respectively.

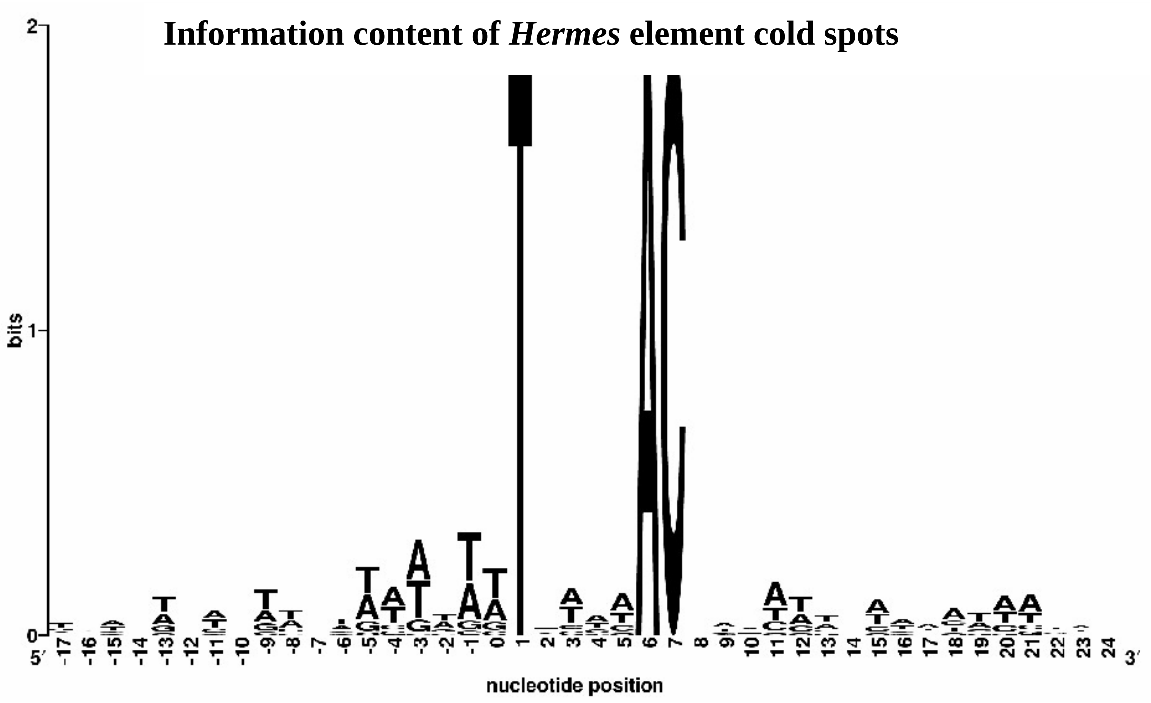


Figure VI-25: Sequence logo providing information content of the cold spots in pGDV1 for the *Hermes* element. Similar to the hot spots, the potential target site is at base 0 and shows the 8bp NTNNNNAC consensus sequence that was used to create this dataset. The cold spots show a preference for A/T nucleotides at positions -3 and -1, but no preference at position -2, which is the opposite trend as seen in the information content of the insertion sites for the *Hermes* element. There is an indication of A/T nucleotides being conserved at position -1, which on the contrary is not seen with the insertion sites. It seems possible that the presence of these conserved A/T nucleotides immediately before the start of the consensus sequence may eliminate the use of these cold spots for integration of the *Hermes* element.

6.4 Profiles of *piggyBac* element insertions in the pGDV1 target plasmid

The second dataset investigated using plasmid based transposition assays were for *piggyBac* element insertions. The bendability profile shown in Figure VI-26 depicts very rapid changes in trends throughout the 200 nucleotide positions. No clear variation in pattern was observed for either the insertion sites or the cold spots at and around the TTAA target sequences. The ranges of bendability values for the insertion sites were between –0.07489 to -0.0094836. The cold spots were more variable in values than the insertion sites, ranging from -0.1176364 to +0.025.

The nucleosome positioning profile was as variable in range as the bendability profile (Figure VI-27). Both insertion sites and cold spots exhibited several changes from being flexible to rigid and vice versa all throughout the 200 bps that were examined. No significant peaks were seen close to the TTAA sequences in either case. The range of variability at the insertion sites was between -14.1936249 and -0.278619. The cold spots were even more variable in flexibility/rigidity and values ranged from -22.8292119 to +3.3813543 (Data not shown).

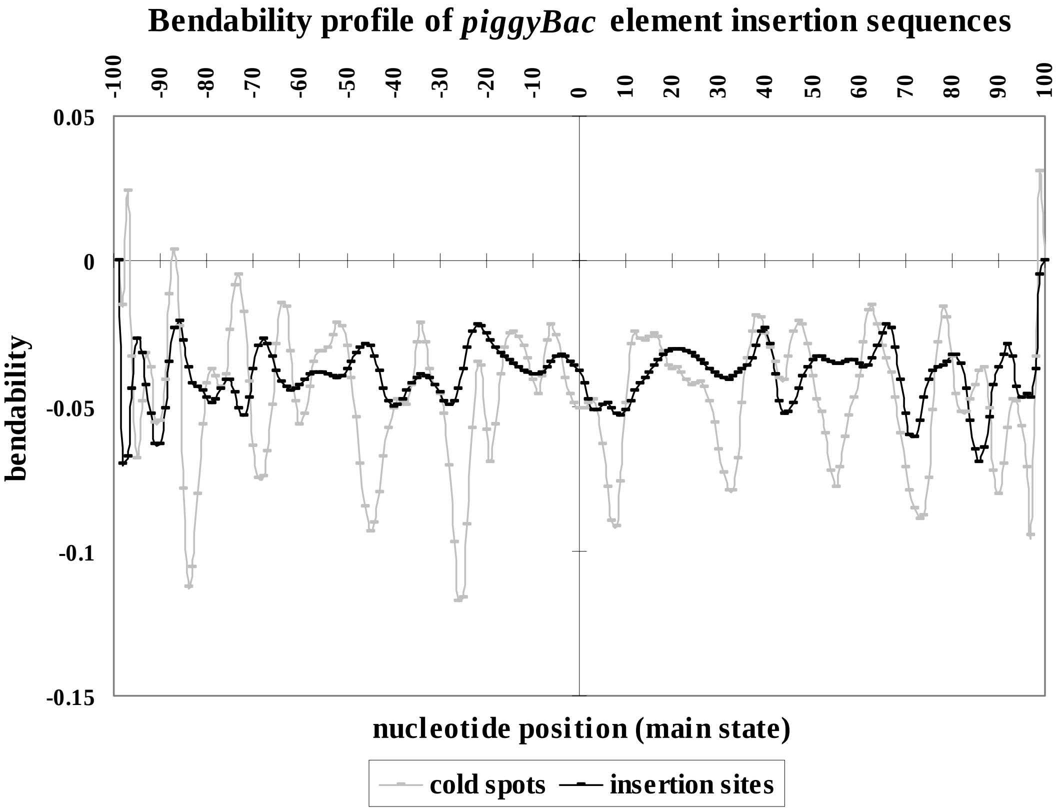


Figure VI-26: DNA bendability profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid. Profile represents a model that analyzed 15 hot spots shown by the black line and 9 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the insertion site at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point. The Y-axis represents the bendability values based on trinucleotide scales ranging from -0.280 (rigid) to +0.194 (very bendable). The less negative or closer the value to zero implies a greater degree of bendability. This profile shows several rapid changes in bendability throughout the entire stretch of 200 bps. The cold spots show positive as well as negative values of bendability. There is no distinct pattern or trend representing a high degree of bendability or rigidity exactly at the insertion point.

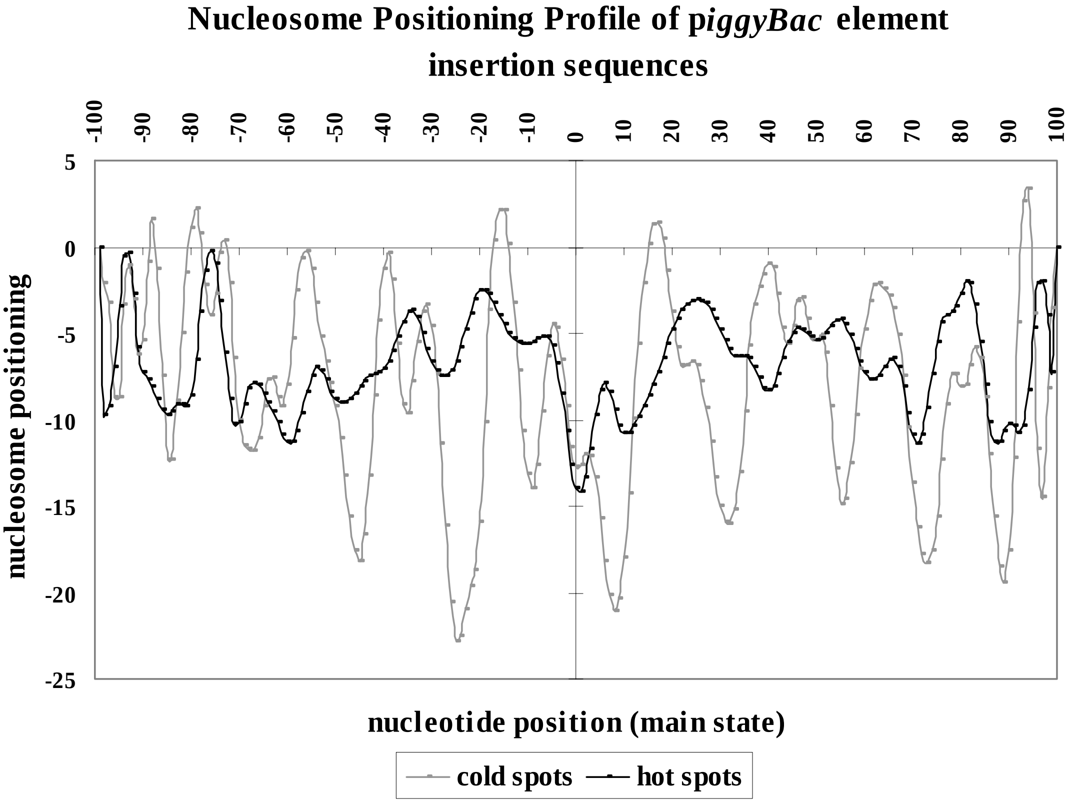


Figure VI-27: DNA nucleosome positioning profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid. Profile represents a model that analyzed 15 hot spots shown by the black line and 9 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the insertion site at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point. The Y-axis represents the nucleosome positioning values based on trinucleotide scales. The more positive or away from zero, the greater the flexibility and vice versa. Similar to the *Tc1* profile, this profile also shows several rapid changes in flexibility throughout the entire stretch of 200 bps. There is no distinct pattern or trend representing a high degree of flexibility or rigidity exactly at the insertion point.

The unsigned nucleosome positioning profile is shown in Figure VI-28. Similar to the signed nucleosome positioning profile, both insertion sites and cold spots exhibited several changes from being flexible to rigid and vice versa all throughout the 200 bps that were examined. The range of variability at the insertion sites was between

13.86945 and 18.66656 and that of the cold spots was from 12.65518 to 24.5126 (data not shown).

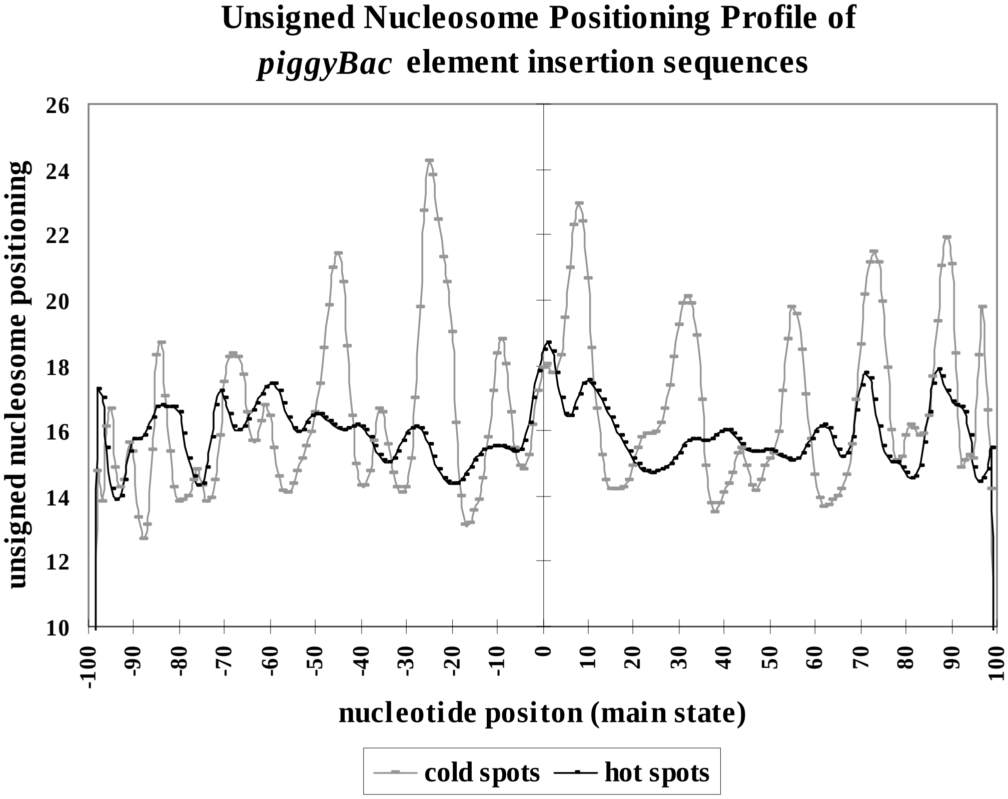


Figure VI-28: DNA unsigned nucleosome positioning profile for *piggyBac* element insertion sequences in the pGDV1 target plasmid. The profile represents a model that analyzed 15 hot spots shown by the black line and 9 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the insertion site at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point. The Y-axis represents the unsigned nucleosome positioning values based on trinucleotide scales. Values closer to zero implies greater degree of flexibility. This profile is consistent with the signed nucleosome positioning profile and does not reveal any coherent pattern of change in trends from before, at and after the insertion site.

Figure VI-29 shows the propeller twist profile of the *piggyBac* element. This profile is again comparable to the previously described bendability and nucleosome positioning profiles of the *piggyBac* element. As seen before, the profiles are random and show fast changing trends varying between regions of flexibility and rigidity. The signal from the cold spots is more varying than that of the insertion sites. The variation in parameter values for the cold spots ranged from -12.3522 to -16.2224, whereas the variation for the insertion sites was between -13.2182 to -14.9417. However, it is seen that the target site at base 0 exhibits one of the highest twist values of -14.9417 indicating rigidity.

Figure VI-30 shows the stacking energy profile of the *piggyBac* element. Both the cold spots and the insertion sites reveal a random pattern all throughout the 200 bps. However, in the case of the *piggyBac* element both the insertion sites and the cold spots reveal a region of instability at the target site.

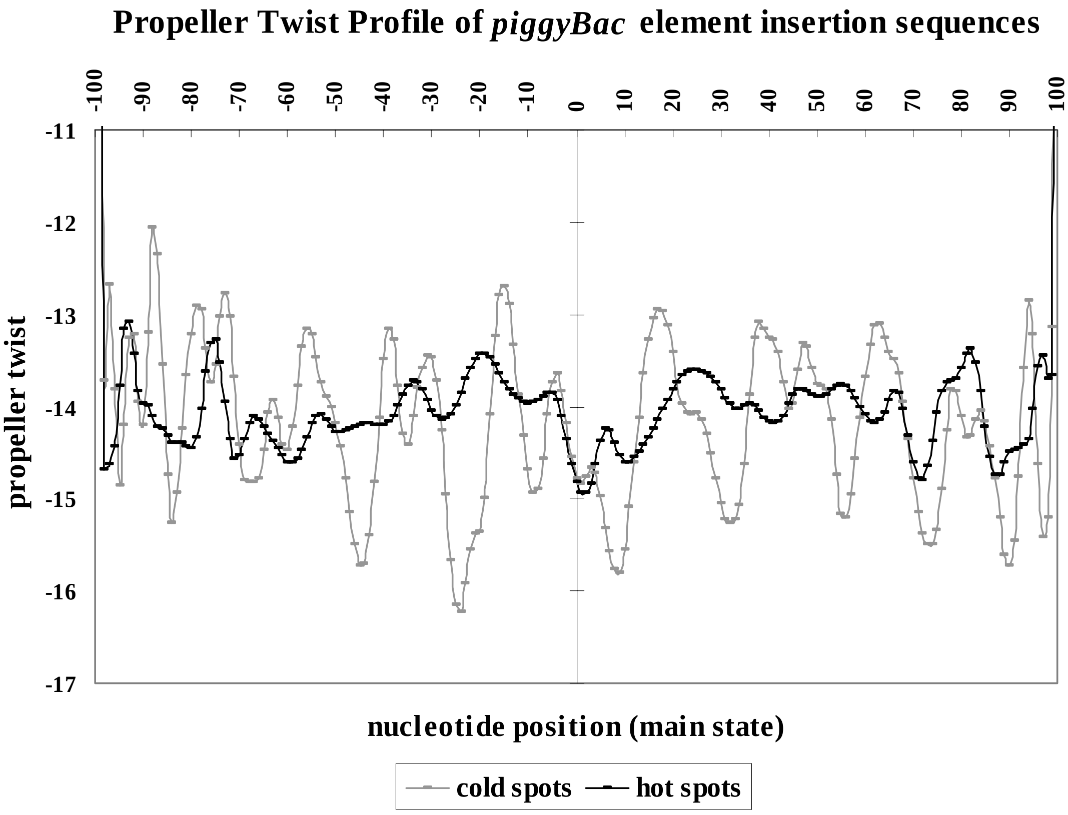


Figure VI-29: DNA propeller twist for *piggyBac* element insertion sequences in the pGDV1 target plasmid. The profile represents a model that analyzed 15 hot spots shown by the black line and 9 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the insertion site at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point. The Y-axis represents the propeller twist values. Flexibility is inferred from less negative values whereas rigidity is implied when values are more negative. The signal from the insertion sites reveals rapid changes in trends from being bendable to rigid and vice versa, but a higher degree of variability is seen with the cold spots. However, the signal at the target site of the insertion sites indicates the highest propeller twist value and corresponds to rigidity.

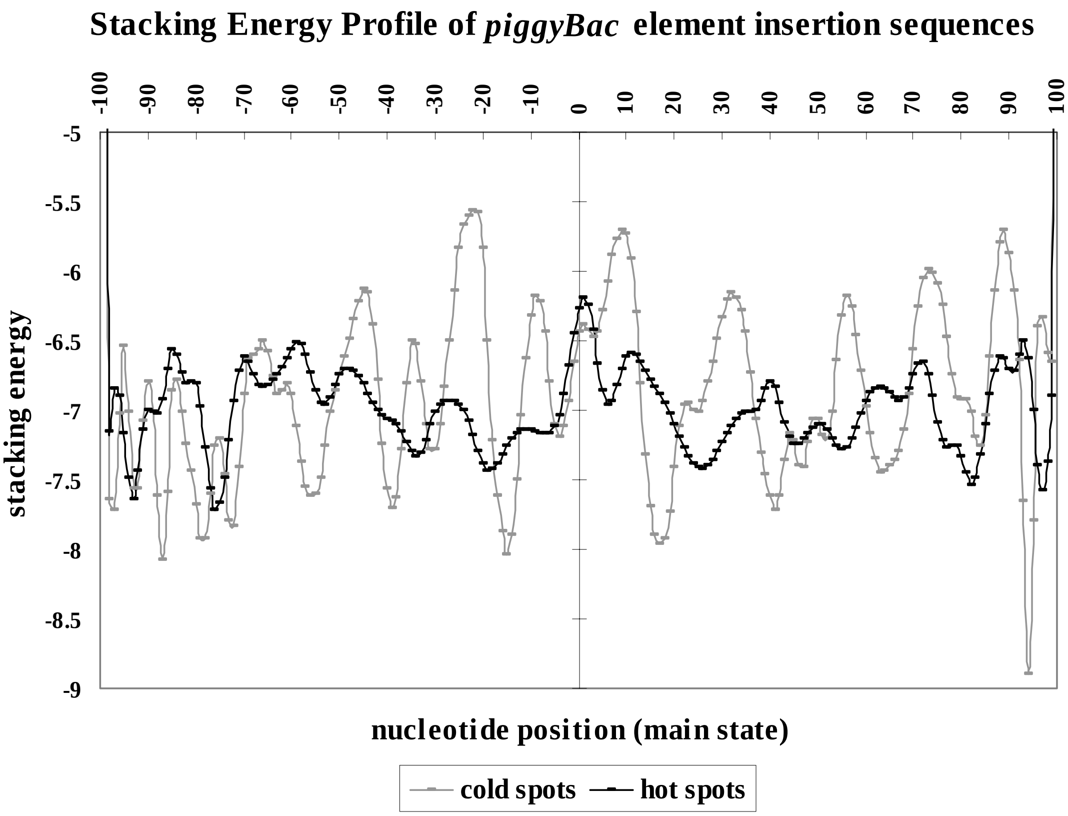
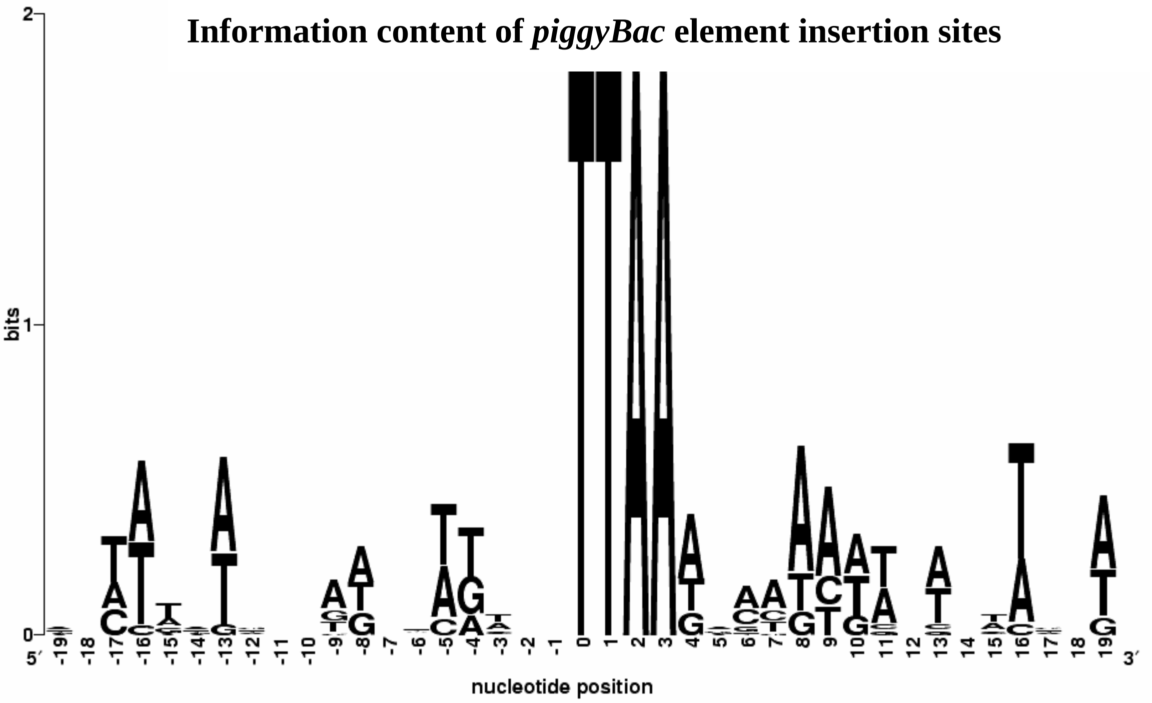


Figure VI-30: DNA stacking energy for *piggyBac* element insertion sequences in the pGDV1 target plasmid. The profile represents a model that analyzed 15 hot spots shown by the black line and 9 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps with the start of the insertion site at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point. The Y-axis represents the stacking energy values. Stability is inferred from more negative values whereas rigidity is implied when values are less negative. The signal from the *piggyBac* element insertion sites reveals rapid changes in trends from being bendable to rigid and vice versa. However, the lowest value for stacking energy is seen close to the target site of the insertion sites and this indicates regional instability.

Information content of the *piggyBac* element insertion sites is shown in figure VI-31 and that of the cold spots is seen in figure VI-32. Both figures reveal the consensus tetranucleotide TTAA sequence between the start of the target site (position 0) to the end of the target site (position 3). This clearly reveals that the *piggyBac* element only recognizes and inserts at TTAA target sites. 19 bps of downstream sequence (positions -1 to -19) and 16 bps of upstream sequence (positions +4 to +19) are also shown in both figures.

The insertion sites reveal no preference for any nucleotide at the 1st (position -1) and 2nd positions (position -2) just before the start of the insertion site. Similarly there is no preference at positions -6, -7, -10 and -11. There is a likelihood of preference for A/T nucleotides at positions -4 and -5. It appears that the upstream flanking sequences keep shifting between positions that exhibit a strong preference for A/T nucleotides and those that do no have any preference at all. However, the downstream sequences seem quite A/T rich which seems likely since the entire pGDV1 target plasmid is approximately 69% AT rich.

The cold spots reveal some similarities and differences in the information content compared with the insertion sites. The first difference between the insertion sites and the cold spots lie in positions -1 and -2 wherein the cold spots show a likely preference for A/T nucleotides at those positions, whereas the insertion sites show the opposite trend in which there is no preference for any nucleotide at those positions. The second difference lies in the cold spots showing a preference for only A/T nucleotides and no other nucleotide at position -7 while in the profile of the insertion sites there is no preference for any nucleotide at the same position. Further in the upstream sequence, position +12 shows no preference for any nucleotide in both profiles, however information at positions +13 and +14 are reversed in the two profiles.



Information content of *piggyBac* element insertion sites

Figure VI-31: Sequence logo providing information content of the insertion sites in pGDV1 for the *piggyBac* element. The potential target site is at base 0 and shows the 4bp TTAA consensus sequence that was used to create this dataset. The insertion sites show preference for A/T nucleotides at positions -4 and -5, but no preference at positions -1, -2, -6, -7, -10 and -11. The downstream sequences seem to be fairly A/T rich and show higher preferences for A/T nucleotides at most positions. No preference for any nucleotide is seen at positions +12 and +14.

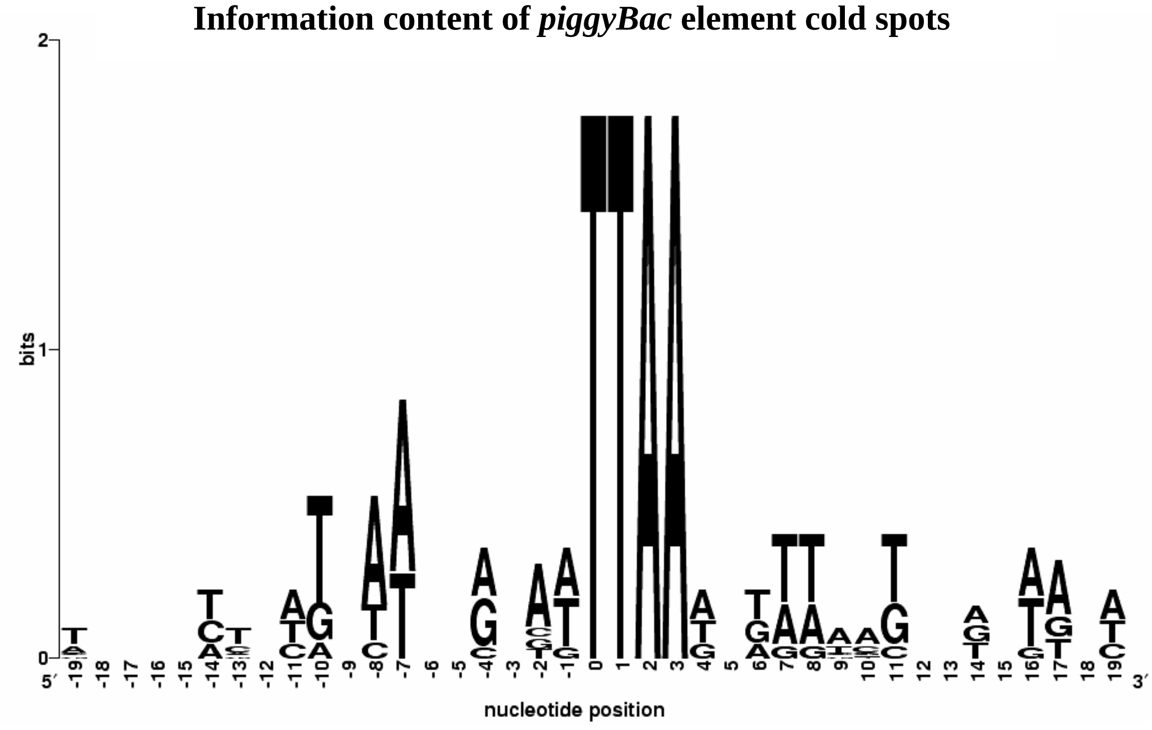


Figure VI-32: Sequence logo providing information content of the cold spots in pGDV1 for the *piggyBac* element. The start of the tetranucleotide TTAA target site is at base 0 and ends at base +3. This shows that the 4bp TTAA consensus sequence was used to create this dataset. The cold spots show a likely preference for A/T nucleotides at positions -1 and -2, whereas the insertion sites show the opposite trend in which there is no preference at all at those positions. In this profile, position -7 shows a preference for only A/T nucleotides and no other nucleotide while in the profile of the insertion sites; there is absolutely no preference for any nucleotide at the same position. These differences in information content in the profiles of the insertion sites and cold spots may be a contributing factor in the selection of certain sites over others.

6.5 Profiles of *Mos1* element insertion sequences in the pGDV1 target plasmid

The final dataset examined based on the plasmid based transposition assays was for the *Mos1* element. Similar to the *piggyBac* and *Tc1* elements, the *Mos1* element did not show any noticeable changes in trends at the insertion point alone, but instead was variable throughout the 200 bps that were analyzed.

Figure VI-33 illustrates the bendability profile of the *MosI* element. The variability in bendability observed was between -0.0457681 and -0.0204143 for the insertion sites, while for the cold spots the ranges were between -0.0457553 and -0.0262451 (data not shown). From the values of bendability, it is seen that the variation in both signals is quite similar and nothing conclusive could be determined about the bendability of the insertion sites and the cold spots at the target site due to the rapidly varying signals across the complete profile.

Similar to the bendability profile, the nucleosome positioning profile shown in Figure VI-34 was also variable and ranged from -11.2379424 to -1.3210879 for the insertion sites and from -9.9924064 to -3.2574979 for the cold spots (data not shown). However, the most negative values corresponding to rigidity were seen at the target site for both insertion sites (-11.2379) as well as cold spots (-9.84614).

The unsigned nucleosome positioning profile is shown in Figure VI-35. As seen in the previous signed nucleosome positioning profile, there is a lot of variability in values of the two signals. The variation for the insertion sites was between 14.08178 and 17.12589 and that of the cold spots was between 15.2671 and 16.80596. Here again for both signals, the highest positive value of unsigned nucleosome positioning was seen at the target site and was approximately 17. This suggests a high degree of regional rigidity in that area.

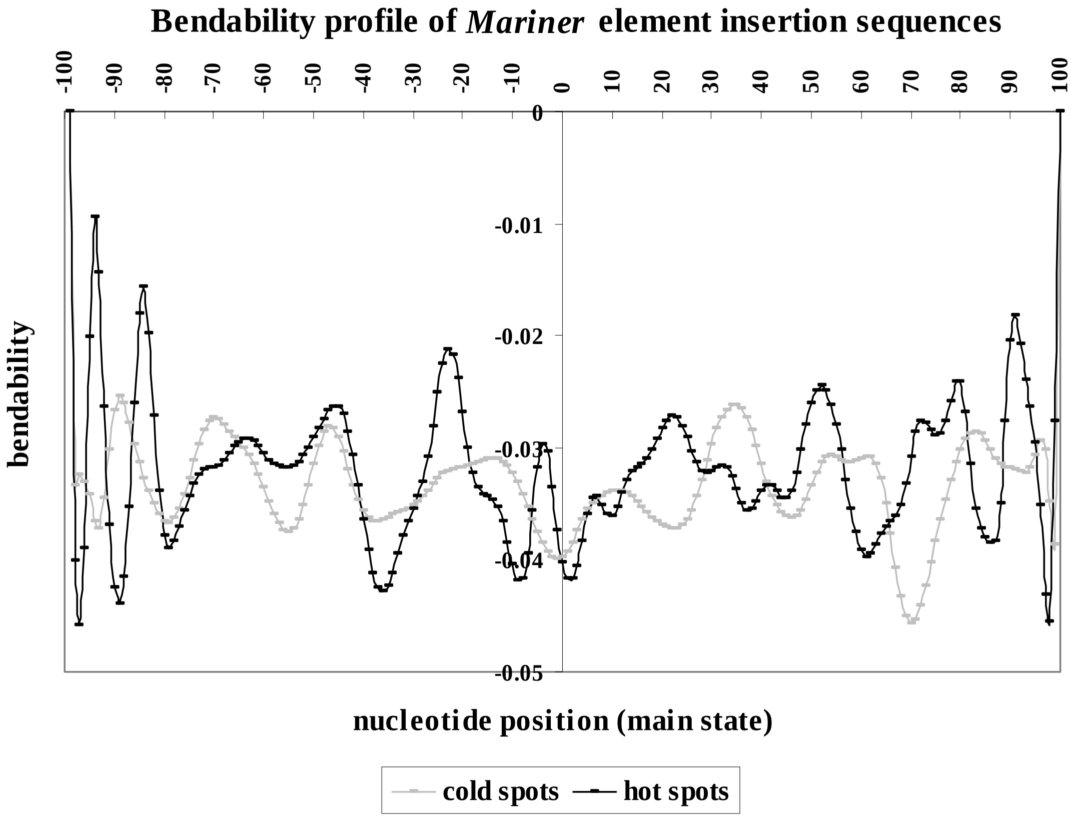


Figure VI-33: DNA bendability for *Mos1* element insertion sequences in the pGDV1 target plasmid. This model analyzed 35 hot spots shown by the black line and 145 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps where the start of the insertion site is at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point are seen on either side. The Y-axis represents the trinucleotide bendability scale. The closer the value is towards zero, the more bendable the DNA is in that region and vice versa. Similar to the *Tc1* and *piggyBac* element bendability profiles, the signal from the *Mos1* element insertion sites and cold spots reveal rapid changes in trends from being bendable to rigid and vice versa. There are no significant changes in trends that are unique to only the region in and around the target site.

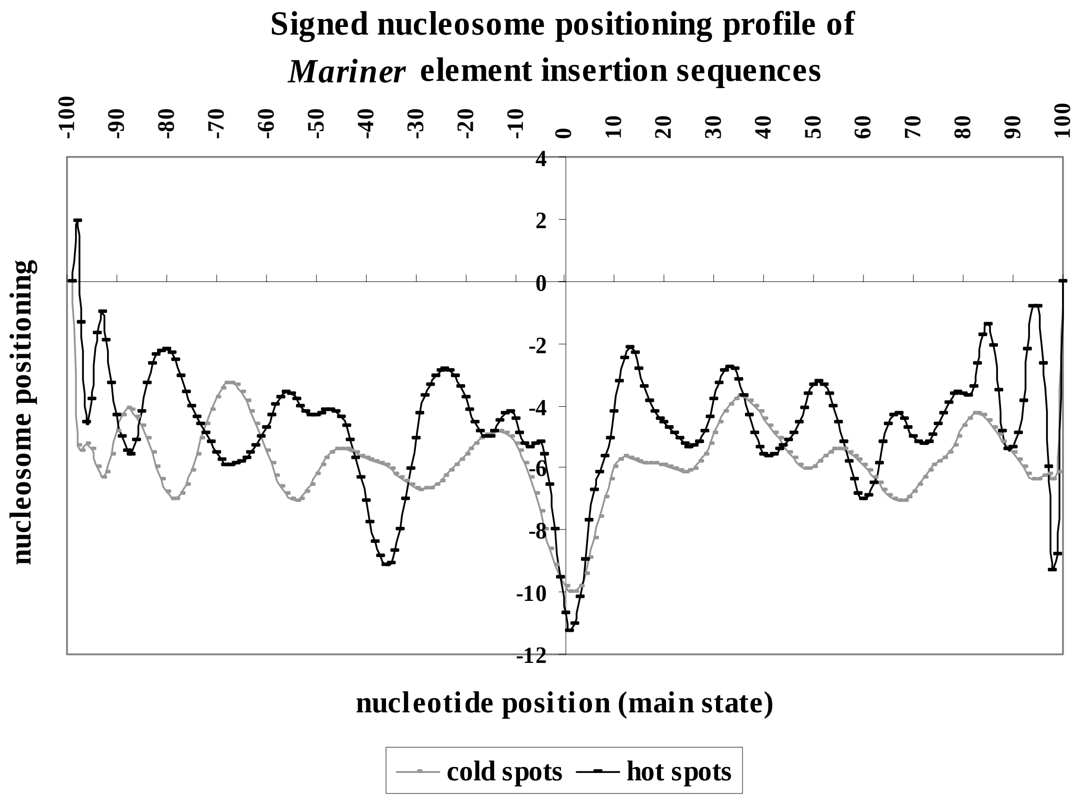


Figure VI-34: DNA nucleosome positioning for *Mos1* element insertion sequences in the pGDV1 target plasmid. This model analyzed 35 hot spots shown by the black line and 145 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps where the start of the insertion site is at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point are seen on either side. The Y-axis represents the trinucleotide nucleosome positioning scale. The higher or more positive the value of nucleosome positioning, the greater the flexibility of DNA in that region. This profile is not as random as the bendability profile. However, it appears as though the trend of the cold spots was almost following that of the insertion sites. The most negative value of nucleosome positioning was seen very close to the target site for both insertion sites (-11.2379) and cold spots (-10.0017). Hence the most rigidity is seen at the target site.

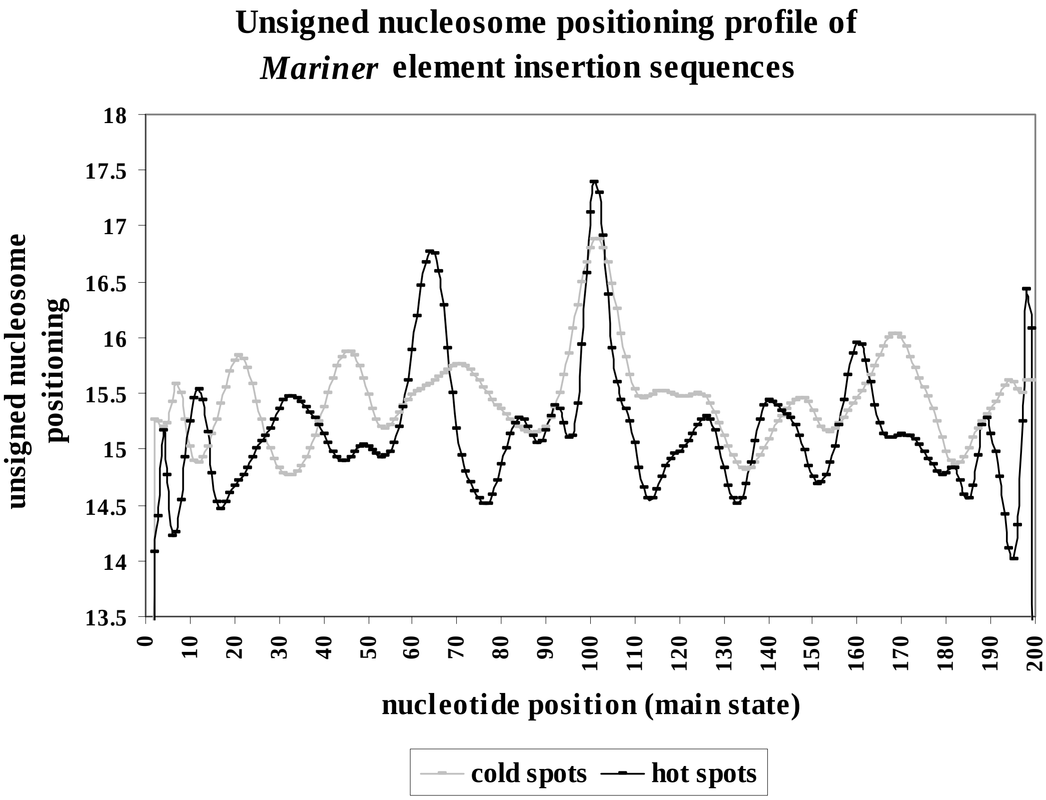


Figure VI-35: DNA unsigned nucleosome positioning profile for *Mos1* element insertion sequences in the pGDV1 target plasmid. This model analyzed 35 hot spots shown by the black line and 145 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps where the start of the insertion site is at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point are seen on either side. The Y-axis represents the unsigned nucleosome positioning values based on trinucleotide scales where values closer to zero implies a greater degree of flexibility and more positive values suggest rigidity. Similar to the signed positioning profile, a huge variation in signal is seen. The variation for the insertion sites was between 14.08178 and 17.12589 and that of the cold spots was between 15.2671 and 16.80596. However the area furthermost away from zero was at the target site and this suggest a highly rigid area at the target site.

The *Mos1* propeller twist profile is seen in Figure VI-36. This profile is in keeping with the other three profiles generated for the *Mos1*element. The highest twist angles corresponding to highest rigidity are seen very near to the insertion site. The variation in the insertion sites were between -12.7503 and -14.6145 while the variation in the cold spots was between -13.6663 and -14.4361.

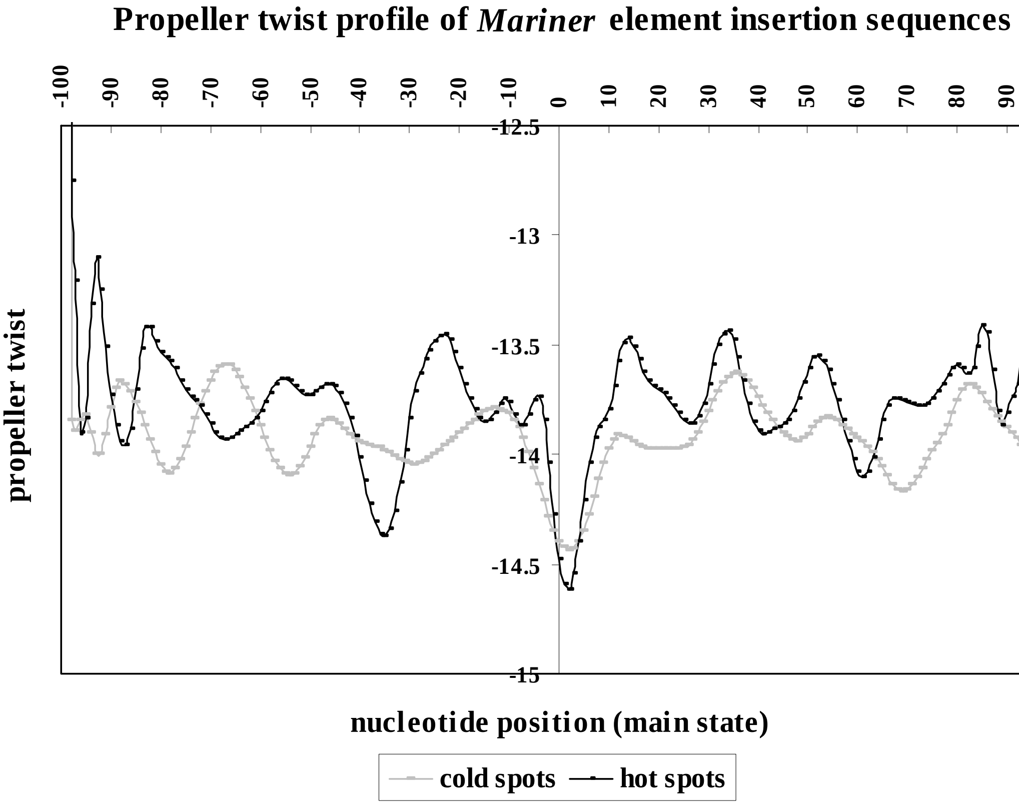


Figure VI-36: DNA propeller twist profile for *Mos1* element insertion sequences in the pGDV1 target plasmid. This model analyzed 35 hot spots shown by the black line and 145 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps where the start of the insertion site is at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point are seen on either side. The Y-axis is a measure of the propeller twist angles which is directly related to rigidity and a higher twist angle refers to higher rigidity. The highest twist angle is seen at the target site and has a value of -14.5927 for the insertion sites while the cold spots have a twist of about -14.425. Both signals appear to suggest rigidity at the insertion site.

The *Mos1* stacking energy profile is seen in Figure VI-37. This profile suggests that the lowest stacking energy values corresponding to highest rigidity are seen very near to the insertion site. The variation in the insertion sites were between -12.7503 and -14.6145 while the variation in the cold spots was between -13.6663 and -14.4361.

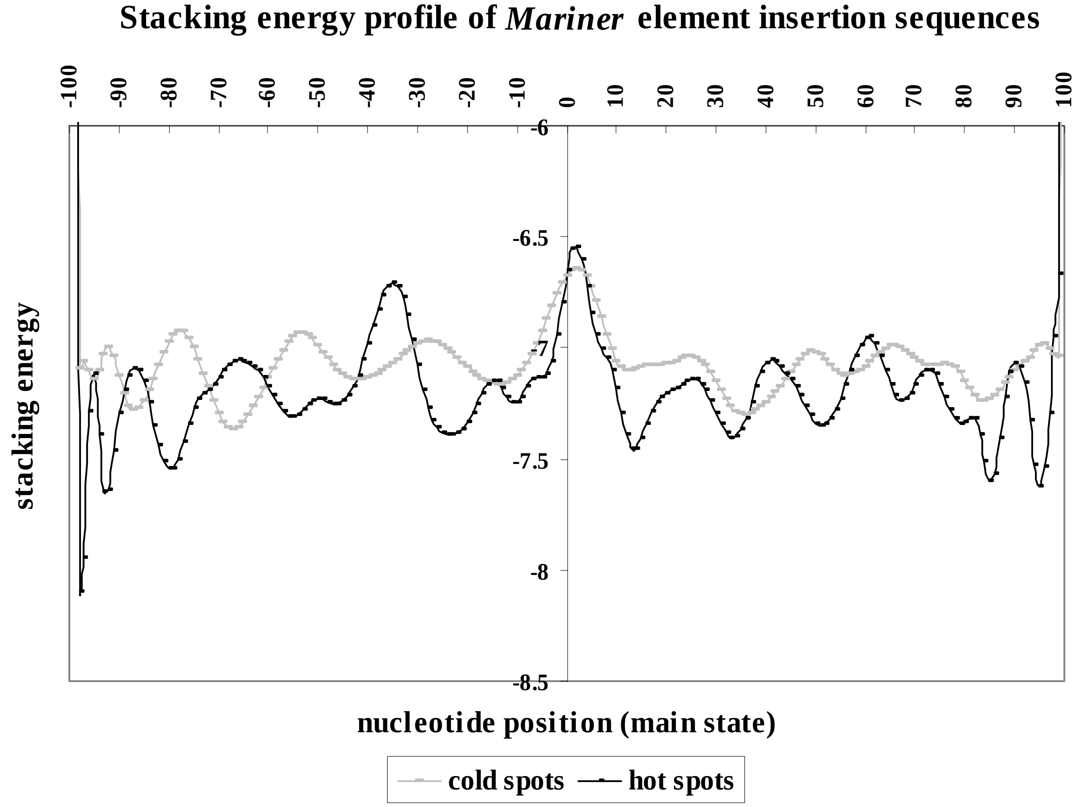


Figure VI-37: DNA stacking energy profile for *Mos1* element insertion sequences in the pGDV1 target plasmid. This model analyzed 35 hot spots shown by the black line and 145 cold spots shown in gray. The X-axis represents nucleotide position ranging from -100 to +100 bps where the start of the insertion site is at base 0 and 100 bps of flanking DNA upstream and downstream of the insertion point are seen on either side. The Y-axis gives a measure of stacking energy where more negative values are associated with higher stability and vice versa. The overall change in trends is seen to be rapidly varying throughout the profile for both insertion sites and cold spots. However, the lowest stacking energy corresponding to instability is observed near the target site and insertion sites have a value of -6.54327 while the cold spots have a stacking energy of about -6.64041.

Information content of the *Mos1* element insertion sites is shown in figure VI-38 and that of the cold spots is seen in Figure VI-39. In both these figures the TA target consensus is conserved at positions 0 and 1. Hence the start of the target site is marked at position 0. For the insertion sites, there seems to be a lack of preference for T/A nucleotides at positions -5, +5 and +6. However, the weblogo profile of the cold spots is consistent in its upstream and downstream flanking sequence and exhibits an overall equal preference for all nucleotides. Hence it is possible that the information content at positions -5, +5 and +6 may serve to distinguish insertion sites from those unused sites (cold spots).

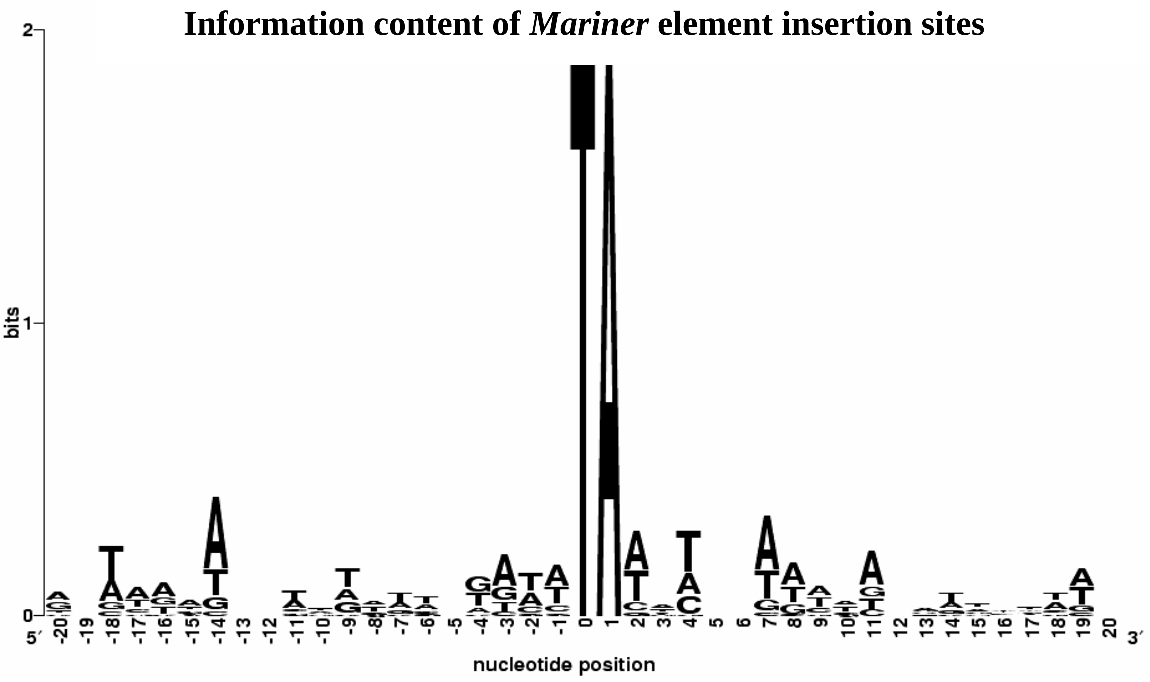


Figure VI-38: Sequence logo providing information content of the insertion sites in pGDV1 for the *Mos1* element. The start of the consensus dinucleotide TA target site is at base 0 and ends at base +1. 20 bps of downstream sequence (positions -1 to -20) and 18 bps of upstream sequence (positions +2 to +20) are shown in this profile. This profile indicates no preference for any nucleotide at positions -5, -12, -13, +5, +6 and +12.

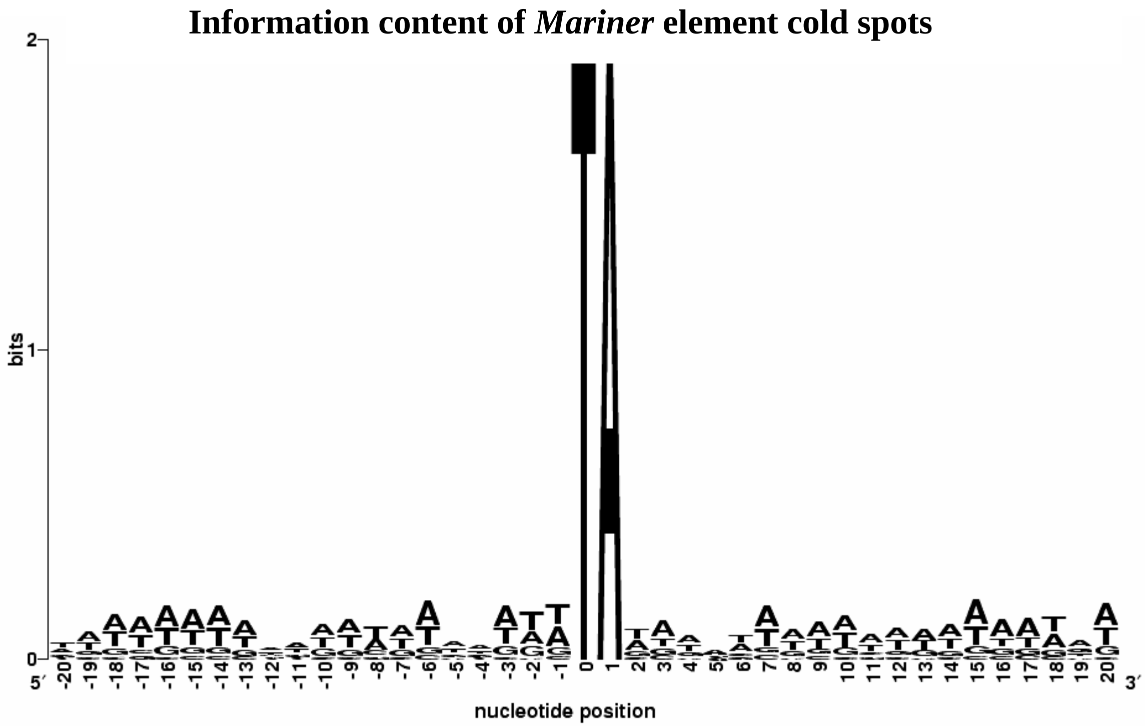


Figure VI-39: Sequence logo providing information content of the cold spots in pGDV1 for the *Mos1* element. The start of the consensus dinucleotide TA target site is at base 0 and ends at base +1. 20 bps of downstream sequence (positions -1 to -20) and 18 bps of upstream sequence (positions +2 to +20) are shown in this profile. It is observed that the overall information content of the flanking sequences do not reveal any particular preference for a nucleotide at one position although there is a high AT content in the flanking sequences. Hence the information content at positions -5, +5 and -6 seen in the profile of the *Mos1* insertion sites may be significant in the selection of those insertion sites.

CHAPTER VII

DISCUSSION AND CONCLUSIONS

This analysis suggests that for certain elements such as the *P* element from *Drosophila melanogaster* and the *Hermes* element from *Musca domestica*, secondary DNA structure plays an important role in target site selection, whereas for the other elements such as the *Tc1* element from *Caenorhabditis elegans*, the *MosI mariner* element from *Drosophila mauritiana* and the *piggyBac* element from *Trichoplusia ni* it appears that secondary structure around target sites may not be critical and that the selection of target sites for these elements may be due to other contributing factors.

7.1 Analysis of *P* element insertions in the *Drosophila melanogaster* genome

It was previously found by Liao *et al*., [10] that the *P* element prefers to integrate into areas of bent DNA. In their analysis of *P* element insertions, a dataset containing 467 insertions was used and each insertion was flanked on either side by at least 250 bps. They examined 12 different DNA parameters and obtained a significant signal at the *P* element insertion site whereas they did not obtain any such signal with the randomly generated sequences.

In this analysis, a software package called HMMpro was used and five DNA function parameters were selected to confirm these findings and hence the results generated in this study were consistent with previous findings. In the previous study, it was observed that the bendability signal was symmetrical around the insertion point.[10] In this study, it was found that amongst all the elements that were analyzed, the flanking sequences around the *P* element insertion point showed the most symmetrical signal of bendability. *P* element insertion sites were previously shown to have a high GC content and triplets CAG. CTG, GAC, GCC, GGC and GTC associated with high values of bendability were predominant.[10]

The weblogo results revealed similar information content for the consensus sequence. In addition, the weblogo profile generated in this study revealed an information pattern unique to the nucleotide positions immediately before the start and following the end of the 8 bp target consensus sequence. It was found that there is a high likelihood of occurrence of an A or a T nucleotide at exactly 3 bps (positions -3 and +10) before the start (position 0) and after the end (position +7) of the consensus sequence. In particular, it is seen that there is a preference for an A nucleotide at position -3 and preference for a T nucleotide at position +10. Positions -2 and -1 immediately preceding the start of the consensus sequence and positions +8 and +9 immediately following the consensus sequence show minimal preference for any nucleotide. Further upstream of position -3 and downstream of position +10, the flanking sequences show no preference for any nucleotide. Hence it is likely that this unique information pattern contained in the 3rd nucleotide positions immediately before and after the consensus sequence perhaps serves to distinguish between target sites that are used as insertion sites and those that are not used. Furthermore, it is possible to suggest that the *P* element target site consensus sequence could be extended to be ANNGGCCAGACNNT with the middle 8bp being duplicated upon insertion.

7.2 Analysis of the *Tc1* element insertions in the *C. elegans* genome

Having confirmed the utility of the HMMPro software and that the *P* element insertions are dependent on both primary sequence and secondary DNA structure, another transposable element, the *Tc1* element from *Caenorhabditis elegans* was then analyzed. The first dataset for *Tc1* element insertion sites consisted of 821 sequences while the second dataset consisted of 22 insertions into the gpa-2 gene of the *C.elegans* genome.

Unlike the *P* element insertion site analysis, no distinct pattern of bendability at the insertion site was observed for the *Tc1* element. Neither were there any noticeable trends in the nucleosome positioning, propeller twist and the stacking energy profiles. Recently, Vigdal *et al.*, [11] studied the structural dependence of *Sleeping Beauty (SB)*, a *Tc1*/*mariner*-type transposable element, which inserts at TAs, for the selection of target sites for integration. They found significant differences in structure between insertion sites and random DNA. They went on to examine the bendability of DNA flanking integration sites and indicated that *SB* prefers to integrate into bent DNA. From their results on bendability, it was observed that at the integration site, the highest value of bendability at *SB* insertions site was approximately 0.05.[11] In this study, the results for the *Tc1* element of *C.elegans* revealed a maximum absolute value of 0.03 for the first dataset and an absolute value of 0.1453026 for the second dataset.

Analysis with *SB* revealed a prominent peak at the insertion site, suggesting high bendability and clearly showed an upward trend going from 0 to 0.05.[11] However, the *Tc1* bendability profile created for the first dataset had very little variation (-0.023 to

-0.032), whereas the variation in the bendability profile of the second dataset (+0.011 to -0.165) was much larger than even that of *SB*. The amplitude of the signals obtained with the second dataset was larger than those of the first dataset with respect to all the five DNA structural parameters examined in this study. An explanation for the larger amplitude range for the second dataset could lie in the number of sequences used in each of these studies. The first dataset used 821 sequences, whereas the second dataset within the gpa-2 gene used only 22 sequences. It seems reasonable that the positive and negative signal amplitudes for a larger number of sequences are averaged out and hence one would expect to get a smaller signal, whereas, significant differences in positive and negative signal amplitudes for a smaller dataset have a greater effect on the overall signal and hence may account for a larger variation in the range of signal amplitude. Analysis of the flanking sequences surrounding *SB* sites of integration showed that these sequences were AT rich, a similar trend was also observed for sequences flanking *Tc1* TEs.

7.3 Analysis of the *Hermes, Mos1* and *piggyBac* element insertions into the pGDV1 target plasmid

Integrations of the *Hermes*, *Mos1* and *piggyBac* transposable elements into the pGDV1 target plasmid were also examined. Using previously published transposition assay results, target sites that had been hit before were determined and this gave rise to the creation of the datasets for the insertion sites. Next, the pGDV1 plasmid sequence was manually edited and all potential targets that could be used were identified based on previously identified consensus sequences. If primary sequence were the only criteria for the selection of a target for element integration, then transposable elements such as *Tc1* and *Mos1*, which depend upon a TA recognition sequence, should in essence have abundant target sites to select from and integrate into. In the case of the *Mos1* element insertions in the pGDV1 plasmid; only 35 out of 180 potential target sites are selected, leaving 145 unused cold spots. The dataset for the *Hermes* element consisted of 76 insertion sites and 51 cold spots, the cold spots being selected based on the 8 bp NTNNNNAC consensus hot spot sequence. The dataset for the *piggyBac* element, which recognizes TTAA sequences, was the smallest and consisted of 15 insertion sites and 9 cold spots. Given the number of cold spots for each element is relatively high, it appears likely that other factors besides primary sequence contribute to target site selection or exclusion. It was hypothesized that it was probable that secondary structure might influence target site selection of these elements in a manner similar to the *P* element. This analysis revealed strong secondary structure dependency for the *Hermes* element and limited importance for the *Mos1* and *piggyBac* elements.

Of these three elements only the *Hermes* element showed distinct changes in trends of bendability and other DNA structural parameters at the insertion point. However, the largest difference in bendability values between the insertion sites and the cold spots of the *Hermes* element was seen immediately upstream of the target consensus sequence. The bendability profiles of the *Mos1* and *piggyBac* elements suggest very rapid transitions in bendability and rigidity throughout the upstream and downstream flanking sequences and do not suggest any pronounced peaks in the near neighborhood of the insertion site alone. The *piggyBac* bendability profile is distinct in that the cold spots exhibit positive values at certain nucleotide positions. This is not the case for the other elements that were examined so far as these exhibited only negative values of bendability. However, the positive values of bendability seen in the cold spots of the *piggyBac* element were observed at the farthest ends of the flanking sequences away from the insertion point and hence may not significantly contribute to target site selection.

Furthermore, the overall ranges in the bendability values of both the insertion sites and the cold spots were between 0 and -0.05 for the *Hermes* and the *Mos1* elements. However, the bendability values of the *piggyBac* element was in a different range altogether between +0.05 to -0.15 comparable only to the ranges as seen in the bendability profile of the second set of *Tc1* insertions within the gpa-2 gene. This implies that the *piggyBac* element had a bendable range that was four times more variable than the *Hermes* and the *Mos1* elements.

The nucleosome positioning profile of the *Hermes* element clearly shows that for the insertion sites there is a gradual change in pattern from being rigid approximately 15 bps upstream from the insertion point to bendable at the insertion point and then rigid again approximately 20 bps downstream from the insertion point. However, the cold spots show exactly the opposite trend wherein the least flexibility is seen at the consensus target sequence, surrounded by regions of flexibility. Contrary to the *Hermes* element, the *Mos1* nucleosome positioning profile of the insertion sites suggests most rigidity at the insertion point. However, one of the features that seem apparent in the

Mos1 nucleosome positioning profile is that the cold spots follow the trend of the insertion sites to a lesser or greater extent throughout the entire profile but especially at the insertion site. Hence both the insertion sites and the cold spots of the *Mos1* nucleosome positioning profile suggest rigidity at the insertion site as opposed to the *Hermes* which reveals opposite trends at the insertion site. For the *piggyBac* element insertions, similar to its bendability profile, the nucleosome positioning profile for the cold spots again showed positive values and ranged from -20 to +2.5, which was the widest amongst all three elements. This variability in the range of values is again comparable only to the nucleosome positioning profile of the *Tc1* insertions of the second set within the gpa-2 gene.

The third parameter considered in this analysis was unsigned nucleosome positioning. Here again only the *Hermes* element exhibited trends that were distinguishable between the insertion sites and the cold spots. There was a significant opposition in the trends of the insertion sites and the cold spots at the insertion site and hence indicated that the insertion sites were much more bendable at the target site than the cold spots. The *Mos1* profile again revealed trends similar to the previous nucleosome positioning profile wherein it appeared as if the cold spots were following the trends of the insertion sites and both signals suggested rigidity at the insertion site. However, the *piggyBac* element insertion sites were the most random and exhibited a wide range of values between 12 and 25 which was almost twice the variation as seen for the *Hermes* and *Mos1* elements.

The fourth DNA structural parameter that was examined was the propeller twist. Here again, the *Hermes* element shows consistency in trends and reveals a pattern of higher bendability of the insertion sites while exhibiting a lower bendable region for the cold spots at the target site. Similar to all the previous profiles, the insertion sites and the cold spots of the *Mos1* element followed each other for most part of the 200 bps. The closest footprints of the two signals were seen at the target site where both indicated rigidity. The *piggyBac* element again exhibited signals that appeared more like noise than a distinct transition in trends. However, the variability of the ranges of values on this profile was not as drastic as the previous profiles. Both the *Hermes* and the *Mos1* elements had a variation of between -12.5 and -15 and the *piggyBac* element kept within close proximity of this range and exhibited values between -11 and -16.5.

The last parameter that was examined in this analysis was the stacking energy. The overall trends that have been established thus far for these three elements were yet again exhibited in the profile of this last DNA structural parameter. The *Hermes* element exhibited trends that were distinguishable between the insertion sites and the cold spots wherein the insertion sites revealed more stability than the cold spots at the target site. A significant opposition in the trends of both signals was apparent at the target site. The *Mos1* profile again revealed trends similar to the previous profiles wherein it appeared as if the cold spots were following the trends of the insertion sites for most part of the profile. Both signals culminated in a peak of almost equal magnitudes at the target site and suggested unstable regions of DNA in that area. The *piggyBac* element was the most random and no distinct trends were identified for either the insertion sites or the cold spots. The overall range of values was between -6.5 and -8 for the *Hermes* and the *Mos1* elements while the *piggyBac* element followed suit with a range of approximately between -5.5 and -9.

Based on variability of trends and relative values of all five DNA structural parameters at the insertion site and in the regions of DNA flanking the consensus sequence, it is seen that for the *Hermes* element insertions, all profiles showed very significant patterns suggesting flexibility at the consensus sequence for the insertion sites. On the contrary, there is evidence to suggest that the DNA is less bendable in the vicinity of the insertion sites of the *piggyBac* and the *Mos1* elements.

The information content of the cold spots and insertion sites revealed by the weblogo results for the *Hermes* element is significant in that it may help understand why some insertion sites show an orientation preference. The profiles of both cold spots and insertion sites show that the 8 bp consensus sequence starts at position 0 and ends at position 7. In the case of the *Hermes* insertion sites, there is no preference for any nucleotide immediately preceding (position -1) or following (position +8), similar to that observed for the *P* element, however, there is an A/T preference at position -2. Analysis of the *Hermes* cold spots detected a preference for A/T nucleotides at positions -3 and -1, but no preference at position -2. The absence or presence of A/T residues at either of these sites could increase or decrease the rate at which *Hermes* targets a potential insertion site. Furthermore, since these differences were observed near only one end of the consensus sequence, this could explain the orientation preference of the *Hermes* element when it is inserting into essentially palindromic 8bp consensus target sites.[4]

Similarly, the information content of the *Mos1* element insertion sites revealed that there was no preference for any nucleotide at positions -5, +5 and +6. The cold spots on the other hand appeared to have an even distribution of A/T nucleotides in both the upstream and downstream sequences. It may be possible that this difference in the information content may serve to target insertions sites preferentially over the cold spots.

For the *piggyBac* element, there were several differences in the information content displayed between the insertion sites and the cold spots. Perhaps the two main differences between the insertion sites and the cold spots lie in upstream and downstream sequences in the close vicinity of the insertion site. The cold spots exhibit a likely preference for A/T nucleotides at bases -1 and -2, whereas the insertion sites show the opposite trend in which there is no preference for any nucleotide at those positions. Additionally, the cold spots exhibit a strong preference for only A or T nucleotides and no other nucleotide at position -7 and on the contrary the insertion sites show no preference for any nucleotide at the same position.

7.4 Conclusions

The dataset for the *piggyBac* element insertions consisted of only 15 insertion sites and 9 cold spots. It is possible that these 9 cold spots could very possibly be preferred sites for integration, but have not yet been hit during transposition assays. Similarly, the *Mos1* dataset consisted of only 35 insertion sites out of a total of 180 potential target sites that could be used for *Mos1* insertions. Of these 35 known insertion sites, only one site has been hit twice and hence could be classified as a hot spot. The absence of previously identified hot spots suggests that researchers have not yet performed an adequate number of transposition experiments in order to saturate all possible insertion sites. The simplest explanation for this observed variability in insertions of transposable elements is that while secondary structural parameters such as bendability influence certain elements in their selection of target sites, other elements may not be as dependent on DNA structure and may have priorities for other factors in addition to secondary structure.

It is also possible that cold spots are negatively influenced by secondary structure so as to prevent the insertion of transposable elements at these consensus target sites, rather than secondary structure necessarily being a positive influence.

Mos1 and *Tc1* insertions occur at TA residues and *piggyBac* insertions occur at TTAA tetranucleotides, while the *Hermes* and *P* elements rely on 8 bp consensus sequences that are highly conserved only at certain nucleotides within the 8 bp consensus. It is possible that primary sequence contributes more significantly for the *Mos1, Tc1* and *piggyBac* elements in selecting insertion sites since their target site recognition sequences are relatively short and invariable. In contrast, the *Hermes* and *P* elements utilize variable 8bp consensus sequences that are not well conserved at all positions. Hence, it seems reasonable to conclude that the *Hermes* and *P* elements may rely on secondary structure as well as primary consensus sequence for the selection of insertions sites and thus explains the structure dependency of the *Hermes* and *P* elements as observed in this study.

A further question to be addressed is whether primary sequences and/or structures act as signals to highlight the presence of a target site, or if they allow easier access or manipulation of the target site for the transposase. The primary sequence need not necessarily act as a signal to indicate a target site, because it is known that certain sites (unused or cold sites) that match the primary consensus sequence are not selected for insertion of TEs. Furthermore, a difference was not found in the local structure between used and unused sites for some TEs, suggesting that particular structures might allow access to potential target sites.

Although there is evidence that certain elements show structure dependency, there is only limited current experimental data to help us verify the hypotheses in this study. In conclusion, from the DNA structure and weblogo profiles for the *P* and *Hermes* elements, it is found that sequences flanking the target sites for these elements have certain structural properties that appear to be important in the integration of the elements into these sites. Furthermore, flanking nucleotides closest to the target sites have significant information content that may also contribute to the selection of these target sites.

Future work on modeling with HMMs could include determining the accuracy of the prediction of the model. In this study all datasets were trained for five iterations which is a good estimate of accuracy within reasonable confidence limits. However, it is possible to train each dataset with different number of iterations in accordance with the size of the dataset and ensure that the model has converged at maxima reflecting the best possible predictions from the model.

REFERENCES

1. Craig, N. L. (1997). Target site selection in transposition. *Annu. Rev. Biochem.* **66**, 437-474.
2. Van Luenen, H. G. & Plasterk, R. H. (1994). Target site choice of the related transposable elements *Tc1* and Tc3 of *Caenorhabditis elegans.* *Nucl. Acids Res.* **22** 262-269.
3. Ketting, R. F., Fischer, S. E. & Plasterk, R. H. (1997). Target choice determinants of the *Tc1* transposon of *Caenorhabditis elegans*. *Nucl. Acids Res.* **25**, 4041-4017.
4. Sarkar, A., Coates, C. J., Whyard, S. & Willhoeft, U. (1997). The *Hermes* element from *Musca domestica* can transpose in four families of cyclorrhaphan flies. *Genetica*, **99**, 15-29.
5. Lewin, B. (2000). *Genes VII*. Oxford University Press, New York.
6. Griffiths, A. J. F., Gelbart, W.M., Miller, J.H. & Lewontin, R.C. (1999). *Modern Genetic Analysis.* W. H. Freeman and Company, New York.
7. McClintock, B. (1949). Mutable loci in maize. *Carnegie Inst. Wash. Yearbook* **48**,142-154.
8. Snustad, D. P. (2003). *Princples of Genetics*. John Wiley & Sons Inc, New York.
9. Prak, E. T. & Kazazian, H. H. Jr. (2000). Mobile elements and the human genome. *Nat. Rev. Genet.* **1**, 134-144.
10. Liao, G., Rehm, E. J. & Rubin, G. M. (2000). Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, **97,** 3347-3351.

11. Vigdal, T. J., Kaufman, C. D., Izsvák, Z., Voytas, D. F. & Ivics, Z. (2002). Common physical properties of DNA affecting target site selection of *Sleeping Beauty* and other *Tc1*/*mariner* transposable elements. *J. Mol. Biol.* **323**, 441-452.
12. Coates, C. J., Turney, C. L., Frommer, M., O'Brochta, D. A. & Atkinson, P. W. (1997). Interplasmid transposition of the *mariner* transposable element in non-drosophilid insects. *Mol Gen Gene,* **253**, 728-733.
13. Lampe, D. J., Grant, T. E. & Robertson, H. M. (1998). Factors affecting transposition of the *Himar1 mariner* transposon in vitro. *Genetics,* **149**, 179-187.
14. Grossman, G. L., Rafferty, C. S., Fraser, M. J. & benedict, M. Q. (2000). The *piggyBac* element is capable of precise excision and transposition in cells and embryos of the mosquito, *Anopheles gambiae. Insect Biochem. Mol. Biol.,* **30**, 909-914.
15. Kuduvalli, P. N., Rao, J. E. & Craig, N. L. (2001). Target DNA structure plays a critical role in *Tn7* transposition. *EMBO J.* **20**, 924-932.
16. O'Hare, K. & Rubin, G. M. (1983). Structures of transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell*, **34**, 25-35.
17. Rio, D.C., Laski, F. A. & Rubin, G.M. (1986). Identification and immunochemical analysis of biologically active *Drosophila P* element transposase. *Cell*, **44,** 21-32.
18. Engels, W. R., Schlitz, D. M., Eggleston, W. B. & Sved, J. (1990). High-frequency *P* element loss in *Drosophila* is homolog-dependent. *Cell,* **62**, 515-525.

19. Kaufman, P. D. & Riao, D. C. (1992). *P* element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. *Cell*. **69**, 27-39.
20. Berg, C. A. & Spradling, A. C. (1991). Studies on the rate and site-specificity of *P* element transposition. *Genetics*. **127**, 515-524.
21. Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T. & Rubin, G. M. (1995). Gene disruptions using *P* transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl. Acad. Sci*., **85**, 10824-10830.
22. Tsubota, S., Ashburner, M. & Schedl, P. (1985). *P* element-induced control mutations at the r gene of *Drosophila melanogaster*. *Mol. Cell. Biol*., **5,** 2567-2574.
23. Kelley, M. R., Kidd, S., Berg, R. L. & Young, M. W. (1987). Restriction of *P* element insertions at the Notch locus of *Drosophila melanogaster*. *Mol. Cell. Biol.,* **7**, 1545-1548.
24. O'Hare, K., Driver, A., McGrath, S. & Johnson-Schlitz, D.M. (1992). Distribution and structure of cloned *P* elements from the *Drosophila melanogaster P* strain '2. *Genet. Res. Camb*, **60**, 33-41.
25. Plasterk, R. H., Izsvak, Z. & Ivics, Z. (1999). Resident aliens: the *Tc1/mariner* superfamily of transposable elements. *Trends. Genet.* **15,** 326-332.
26. Emmons, S.W., Yesner, L., Ruan, K. S. & Katzenberg, D. (1983). Evidence for a transposon in *Caenorhabditis elegans*. *Cell,* **32**, 55-65.
27. Vos, J.C., van Luenen, H. G. & Plasterk, R. H. (1993). Characterization of the *Caenorhabditis elegans Tc1* transposase in vivo and in vitro. *Genes Dev.,* **7,** 1244-1253.

28. Colloms, S.D., van Luenen, H. G. & Plasterk, R. H. (1994). DNA binding activities of the *Caenorhabditis elegans Tc3* transposase. *Nucleic Acids Res*. **22**, 5548-5554.
29. Medhora M, Maruyama K, Hartl DL. (1991). Molecular and functional analysis of the *mariner* mutator element *Mos1* in Drosophila. *Genetics*, **128,** 311-8.
30. van Luenen, H. G., Colloms, S. D. & Plasterk, R. H. (1994). The mechanism of transposition of *Tc3* in *C. Elegans. Cell*, **79**, 293-301.
31. Eide, D. & Anderson, P. (1988). Insertion and excision of *Caenorhabditis elegans* transposable element *Tc1*. *Mol. Cell. Biol.*, **8,** 737-746.
32. Mori, I., Benian, G. M., Moerman, D. G. and Waterston, R. H. (1988). Transposable element *Tc1* of *Caenorhabditis elegans* recognizes specific target sequences for integration. *Proc. Natl. Acad. Sci. USA,* **85**, 861-864.
33. Vos, J. C., De Baere, I. & Plasterk, R. H. (1996). Transposase is the only nematode protein required for in vitro transposition of *Tc1*. *Genes Dev.*, **10**, 755-761.
34. Duret, L., Marais, G. and Biémont, C. (2000). Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans. Genetics*, **156,** 1661-1669.
35. Bryan, G., Garza, D. & Hartl, D. (1990). Insertion and excision of the transposable element *mariner* in *Drosophila*. *Genetics,* **125**, 103-114.
36. Sarkar, A., Yardley, K., Atkinson, P. W., James, A. A. & O'Brochta, D. A. (1997). Transposition of the *Hermes* element in embryos of the vector mosquito, *Aedes aegypti*. *Insect Biochem. Mol. Biol.,* **27**, 359-363.

37. O'Brochta, D. A., Warren, W. D., Saville, K. J. & Atkinson, P. W. (1994). Interplasmid transposition of *Drosophila hobo* elements in non-drosophilid insects. *Mol. Gen. Genet.*, **244**, 9-14.

38. Saville, K. J., Warren, W. D., Atkinson, P. W. & O'Brochta, D. A. (1999). Integration specificity of the *hobo* element of *Drosophila melanogaster* is dependent on sequences flanking the integration site. *Genetica*, **105**, 133-147.

39. Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E. & Fraser, M. J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, **172**, 156-69.

40. Lobo, N., Li, X. & Fraser, J. (1999). Transposition of the *piggyBac* element in embryos of *Drosophila melanogaster*, *Aedes aegypti* and *Trichoplusia ni. Mol. Gen. Genet.*, **261**, 803-810.

41. Baldi, P. & Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach.* Cambridge, MA: MIT Press.

42. HMMPro Website. http://www.netid.com Date: 03/2002.

43. Weblogo Website. http://weblogo.berkeley.edu Date: 03/2002.

44. Schneider, T. D. & Stephens, R. M. (1990). Sequence Logos: A new way to display consensus sequences. *Nucl. Acids Res.*, **18,** 6097-6100.

45. Shaner, M.C., Blair, I. M. & Schneider, T. D. (2000). Sequence Logos: A powerful, yet, simple tool. (1993). *Proceedings of the Twenty-Sixth Annual Hawaii*

International Conference on System Sciences, Volume 1: Architecture and Biotechnology Computing, 813-821.

46. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Sequence-dependent bending propensity of DNA as revealed by DNaseI: parameters for trinucleotides. *EMBO J*, **14**, 1812-1818.
47. Lahm, A. & Suck, D. (1991). DNase I induced DNA conformation - 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.* **222**, 645-667.
48. Suck, D. (1994). DNA-recognition by DNase I. *J. Molec. Recognition,* **7**, 65-70.
49. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dyn.,* **13**, 309-317.
50. Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol*., **191,** 659-675.
51. Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucl. Acids Res.,* **22**, 5497-5503.
52. Calladine, C. R. & Drew, H. R. (1987). Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.,* **192**, 907-918.
53. Pedersen. A., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA Polymerase II promoters. *J. Mol. Biol*., **281**, 663-673.
54. Baldi, P., Brunak, S., Chauvin, Y. & Krogh, A. (1996). Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.*, **263**, 503-510.

55. el Hassan, M. A. & Calladine, C. R. (1996). Propeller twisting of Base pairs and the conformational mobility of dinulceotide steps in DNA. *J. Mol. Biol.*, **259**, 95-103.
56. Ornstein, R. L., Rein, R., Breen, D. L. & Macelroy, R. D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers,* **17**, 2341-2360.
57. Berkeley *Drosophila* Genome Project (BDGP) Website. http://www.fruitfly.org Date: 05/2002.
58. *C. elegans* Genome Project (AceDB) Website. http://www.sanger.ac.uk/ Date: 04/2002.
59. *Tc1* End sequences. http://elegans.swmed.edu/tc1s/ Date: 04/2002.
60. Rorth, P., Szabo, K., Bailey, A., Laverty, T., Rehm, J., Rubin, G. M., Weigmann, K., Milan, M., Benes, V., Ansorge, W. & Cohen, S. M. (1998) Systematic gain-of-function genetics in *Drosophila. Development,* **125**, 1049-1057.
61. Rorth, P. (1996). A modular misexpression screen in *Drosophila* detecting tissue-specific phenotypes. *Proc. Natl. Acad. Sci. USA*, **93**, 12418-12422.
62. Korswagen, H. C., Durbin, R. M., Smits, M. T. & Plasterk, R. H. (1996). Transposon *Tc1*-derived, sequence tagged sites in *Caenorhabditis elegans* as markers for gene mapping. *Proc. Natl. Acad. Sci. USA,* **93**, 14680-14685.

APPENDIX A

```
import java.lang.*;
import java.io.*;

public class positive
{

  private String sequence;
  private int[] position;
  private String temp;

  public positive()
  {
  try
  {
    // The sequence is found in a text file called pgdv1.txt

    DataInputStream fin = new DataInputStream(
                             new FileInputStream("pgdv1.txt"));

    // The output sequence gets the position from position.txt and
    // stores 100 base pairs before and after the position, into the
    // output  file

    DataOutputStream file=new DataOutputStream(new
                         FileOutputStream("pgdv1positive.txt"));
    // The file which contains the different locations is position.txt

    DataInputStream filein = new DataInputStream(new
                             FileInputStream("position.txt"));
    sequence=new String();

    position=new int[1000];
    int i=0;

    // The last line of the input file pgdv1.txt is a @. So if we come
    // accross an @, we stop reading the pgdv1.txt file. The last line
    // of the position file is -1. We stop reading when we come accross
    // a -1.

    try{
      java.lang.StringBuffer temp1 = new java.lang.StringBuffer();
      temp = "";
      temp = fin.readLine();
      while (temp.compareTo("@")!=0)
      {
        temp1.append(temp);
        temp = fin.readLine();
      }
      sequence = temp1.toString();
      int pos=0;
      pos = filein.readInt();
      while (pos !=-1)
      {
        position[i] = pos;
        i++;
        pos = filein.readInt();
      }
    }catch(java.io.IOException fnfe)


    for (int j=0;j<i;j++)
    {
      java.lang.StringBuffer temp1=new StringBuffer();
      temp=new String("");
      int x=position[j];
      int k=j+1;
      // If x is less than 100 or greater than 2477, then we wrap
      // around the sequence.
      if (x<=100)
      {
        temp1.append(">Hermes("+k+") inserted at base "+x+" +");
        int wrap=100-x;
        String ttemp=new String("");
        ttemp=sequence.substring(0,x+101);
        String ttttemp=sequence.substring(sequence.length()-wrap-
                                          1,sequence.length());
        temp=ttttemp+""+ttemp;
      }
      else if (x>2477)
      {
        temp1.append(">Hermes("+k+") inserted at base "+x+" +");
        int wrap=100-(sequence.length()-x);
        String ttemp=new String("");
        ttemp=sequence.substring(x-101,sequence.length());
        String tttemp=sequence.substring(0,wrap);
        temp=ttemp+""+tttemp;
      }
      else
      {
        temp=sequence.substring(x-101,x+100);
        System.out.println(temp.length());
        temp1.append(">Hermes("+k+") inserted at base "+x+" +");

      }

      for (int y=0;y<4;y++)
      {
        temp1.append("\n"+temp.substring(50*y,50*(y+1)));
      }
      temp1.append(temp.charAt(200)+"\n");
      file.writeBytes(temp1.toString());
```

```
    }
  }catch(java.io.IOException fnfe){}

  }

  public static void main(String[] args )
  {
    positive plus=new positive();
  }

}
```

APPENDIX B

```
import java.lang.*;
import java.io.*;
import java.lang.String;

public class negative
{

  private String sequence;

  public negative()
  {
    String temp;
    sequence=new String("");
    int[] position=new int[1000];
    int i=0;

    try
    {

      DataInputStream filein=new DataInputStream(
                          new FileInputStream("hermescold.txt"));



      // The last line of the input file is a @. So if we come accross
an
      // @, we stop reading the file. The last line of the position
file
      // is -1. We stop reading when we come accross a -1.


        java.lang.StringBuffer temp1 = new java.lang.StringBuffer();
        temp = "";
        temp = filein.readLine();
        while (temp.compareTo("@")!=0)
        {
          temp1.append(temp);
          temp = filein.readLine();

        }
        temp1.reverse();
        sequence = temp1.toString();
        int pos=0;
        pos = filein.readInt();
        while (pos !=-1)
        {
          position[i] = pos;
          i++;
          pos = filein.readInt();
        }
    }catch(java.io.IOException fnfe) {}
    java.lang.StringBuffer sequence1= new
java.lang.StringBuffer(sequence);
    for (int h=0;h<sequence.length();h++)
    {
       if (sequence.charAt(h)=='A' )
       {
         sequence1.replace(h,h+1,"T");
         continue;
       }
       if (sequence.charAt(h)=='T')
       {
         sequence1.replace(h,h+1,"A");
         continue;
       }
       if (sequence.charAt(h)=='C')
       {
         sequence1.replace(h,h+1,"G");
         continue;
       }
       if (sequence.charAt(h)=='G')
       {
         sequence1.replace(h,h+1,"C");
         continue;
       }
    }
    sequence = sequence1.toString();
    try {
    DataOutputStream file=new DataOutputStream(
                            new
FileOutputStream("outputHermescold.txt"));
    for (int j=0;j<i;j++)
    {
      java.lang.StringBuffer temp1=new StringBuffer();
      temp=new String("");
      int x=position[j];
      int k=j+1;
      if (x<=100)
      {
        temp1.append(">Hermes Cold Spots("+k+") inserted at base "+x+"
+");
        int wrap=100-x;
        String ttemp=new String("");
        ttemp=sequence.substring(0,x+101);
        String ttttemp=sequence.substring(sequence.length()-wrap-1,
                                          sequence.length());
        temp=ttttemp+""+ttemp;
      }
      else if (x>2477)
      {
        temp1.append(">Hermes Cold Spots("+k+") inserted at base "+x+"
+");
        int wrap=100-(sequence.length()-x);
        String ttemp=new String("");
        ttemp=sequence.substring(x-101,sequence.length());
        String tttemp=sequence.substring(0,wrap);
        temp=ttemp+""+tttemp;
      }
      else
      {
        temp=sequence.substring(x-101,x+100);
        System.out.println(temp.length());
        temp1.append(">Hermes Cole Spots("+k+") inserted at base "+x+"
+");

      }

      for (int y=0;y<4;y++)
      {
        temp1.append("\n"+temp.substring(50*y,50*(y+1)));
      }
      temp1.append(temp.charAt(200)+"\n");
      file.writeBytes(temp1.toString());

    }
    }catch(java.io.IOException ioe){}
  }

  public static void main(String[] args )
  {
    negative minus=new negative();
  }

}
```

VITA

Name: Andrea Marian Julian

Permanent Address: 801, Spring Loop,
Apt 2403, College Station,
Texas – 77840, USA.

Date of birth: August 21, 1980.

Education: Texas A&M University
College Station, TX – 77843
Master of Science
Biomedical Engineering, 2003.

Easwari Engineering College
Madras University
Chennai, India.
Bachelor of Engineering,
Electronics & Instrumentation Engineering, 2001.