An Investigation of Palindromic Sequences in the *Pseudomonas fluorescens* SBW25 Genome Bachelor of Science Honors Thesis

Lina L. Faller
Department of Computer Science
University of New Hampshire

June 2008

Abstract

Palindromic sequences are commonly found repetitive elements in the genomes of all organisms. While many studies of bacterial genome sequences have noted these features, our knowledge of the origin and function of these elements is still incomplete. In this study, we investigate the evolution of palindromic sequences found in the genome of *Pseudomonas fluorescens* SBW25. Using a collection of computer scripts written in Perl, as well as the Blast software suite, we isolated patterns and investigated their evolutionary behavior throughout the genome. We succeeded in identifying a group of closely related palindromes that are repeated throughout the genome. Based on statistical tests and the rigorous preservation of the sequences, we conclude that this group of patterns may contribute to the coordinated expression of genes in this organism.

1 Introduction

1.1 Background

In the past few decades, much of the research in the field of genetics has begun to focus on analyzing the entire set of instructions, the so-called genome sequences, for an organism. With advancing technology it is now routine to begin the characterization of an organism by sequencing its entire genome, and we have been able to determine the genetic information of many creatures, including humans and thousands of other species as well.

A genome is the complete string of characters that contains the instructions, or genes, for the organism to function and reproduce. It is like a book written without punctuation in a language containing only four letters. For example, the human genome has about three billion letters, the genome of a nematode worm has about one hundred million, and the sequence of the bacterium Escherichia coli has around five million letters. The four letters making up the alphabet of deoxyribonucleic acid (DNA) are the nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C).

The DNA molecule is formed by two chains, or "strands," of nucleotides. Hydrogen bonds among the bases help support this structure that will curl into a double helix. Due to chemical properties of the nucleotides, two hydrogen bonds will usually bind an A and a T molecules together, and three hydrogen bonds can hold G and C molecule together. In this double-stranded scenario, we consider one DNA strand the "forward" strand and the other one the "reverse complement."

1.2 Palindromes

Determining the string of characters is just one step in the process; we need to understand how these characters work to allow the organisms to function and interact with their environment. One way we can explore the function of a DNA sequence is to look for unusual patterns. Here we focus on a special pattern of nucleotides called palindromes.

As defined by the Oxford English Dictionary, a “palindrome” is “a word or sequence of words that reads [. . .] the same backwards as forwards” [4]. Examples of the English language include words such as “madam” or “level.”

For this analysis, we consider a palindromic DNA sequence to be one where reading the forward strand will yield the sequence of the reverse complement strand backwards (see Figure 1).

1.3 Pseudomonas

The genus *Pseudomonas* belongs to the domain of bacteria. The functions of bacteria species range from fixing nitrogen from the atmosphere, treating sewage, producing cheese and yogurt, to causing infectious diseases such as cholera, anthrax, or leprosy.

The *Pseudomonas* genus in particular comprises a variety of different species flourishing in both terrestrial and aquatic habitats. While some specimens are human or plant pathogens, the species we study, *Pseudomonas fluorescens* SBW25, promotes plant growth by colonizing a plant’s roots and leaves [5].

Recently, the complete genome of this bacterium was sequenced (Paul B. Rainey and the Sanger Institute, unpublished) and now the challenge is to use that sequence to better understand what gives this bacterium its special characteristics. We have focused on a search for palindromes with the goal of providing specific functional predictions for DNA sequences.

2 Methods

We wrote computer programs in Perl [6] and Bash that identified palindrome patterns by applying a sliding window analysis over the entire *Pseudomonas* genome. We then used NCBI-BLAST [7] to determine whether or not a specific pattern was repeated in the genome using a percent identity threshold of 90%.

Using this procedure, we collected data on even-length palindrome patterns ranging from 14 to 38 bases in length. Based on the size of the genome (about 6.7 Mb) shorter patterns would not be statistically significant as they could occur by random chance. Applying the algorithms to look for palindromes 40 or more bases long did not provide any data.

To establish a measure of sequence diversity between patterns, we conducted two separate analyses. One investigated perfect palindromes, where the palindromicity of a pattern needs to be conserved. The second analysis examined imperfect palindromes that required 90% or more of the letters to contribute to the palindromic nature of the pattern (see Figure 1).

Finally, we extracted 2000 base pair sequences flanking a group of identical palindrome sequences. We compared these sequences to known genes in related species.

```
5’ -- ATTTTTGCAAAAAT -- 3’ 5’ -- ATTTTTGGAAAAAT -- 3’
3’ -- TAAAAACGTTTTTA -- 5’ 3’ -- TAAAAACCTTTTTA -- 5’
```

Figure 1: A “perfect palindrome” and an “imperfect palindrome” sequence.

3 Results and Discussion

3.1 Perfect and Imperfect Patterns

A first analysis of perfect and imperfect palindromic motifs taken from the original genome sequence is illustrated in Table 1. While there are many perfect palindrome sequences of length 14, the numbers drop drastically as the length of the sequence increases. Compared to the number of perfect palindromes, the number of imperfect palindromes found in the genome levels off more slowly as motif length increases. A sudden drop in numbers occurs at motif length 34. The largest palindromes found overall are 38 bases long.

The palindrome motifs were then blasted back against the genome and all hits that showed a perfect alignment length, i.e. only hits of equal length as the query sequence, and a percent identity score of 90%

Length of Perfect Palindrome Query Sequences	*Number of Perfect Palindrome Motifs Extracted from Genome*	*Number of Imperfect Palindrome Motifs Extracted from Genome*
14	535	10965
16	171	4047
18	54	1283
20	22	4885
22	9	1822
24	5	790
26	4	371
28	4	219
30	3	484
32	2	302
34	2	2
36	1	1
38	1	1
40	0	0

Table 1: Comparing the number of perfect and imperfect palindrome motifs found in the genome.

or greater were filtered out for subsequent analysis. Table 2 outlines the number of total hits returned by Blast using the criteria described above.

A comparison of open reading frame data for the genome and the location of these repeats revealed that perfect palindromes seem to propagate mostly in intergenic regions, while imperfect palindromes do occur within genes. For this analysis, only hits that were located completely between or within genes were counted (see Table 3). A subsequent analysis revealed that some hits partially overlap coding regions (see Table 4). The open reading frame data was provided by the Sanger Institute (unpublished).

3.2 Distance Data

For each motif returning one or more blast hits, a “Distance Score” was calculated by averaging all the percent identity scores and then subtracting this value from 100. Analyzing this data for all the motif hits of a palindrome with a given length shows a trend in sequence diversity (see Figure 2). When graphing the data, two groups seem to emerge among the imperfect palindrome motifs; one group of motifs shows lower numbers of copies and another group shows overwhelmingly large numbers of hits. The perfect palindrome sequences did not follow this trend, as the number of repeats in general were very low (ranging from 2 to 14 repeats per motif).

3.3 Motif Diversity

To evaluate the differences between individual palindrome sequences, a “Difference Score” was calculated by comparing individual bases in a set of same-length motifs. To obtain an average score, this number was then divided by the total number of motifs in the set. The perfect palindromes did not show a great diversity between sequences. The imperfect palindromic sequences, on the other hand, demonstrate great diversity both within and between the two groups identified previously (see Section 3.2 and Figure 2). In Group A, which contains the sequences that repeat between 2 and 25 times throughout the genome, the sequence diversity seems to stay steady around 0.70. The sequences in Group B, on the other hand, show many more differences between sequences, especially among the motifs between 20 and 28 bases long (see Table 5).

3.4 The 90% Identity Score Threshold

When filtering out hits from the initial Blast search, a threshold of an identity score of 90% was used. As with many parameters used to define a data set, this threshold introduces both opportunities for more

Length of Palindrome Query Sequences	*Number of Hits Returned by Blast (Perfect Palindromes)*	*Average Percent Identity Score (Perfect Palindromes)*	*Number of Hits Returned by Blast (Imperfect Palindromes)*	*Average Percent Identity Score (Imerfect Palindromes)*
14	1254	100.00	17411	100.00
16	394	99.27	5552	98.91
18	116	99.62	1693	99.41
20	48	99.58	6075	99.15
22	22	99.17	3022	98.11
24	10	100.00	2186	96.61
26	8	100.00	1725	95.94
28	8	100.00	1348	95.94
30	8	97.50	1486	96.19
32	12	94.79	989	96.08
34	12	95.10	12	95.10
36	10	94.44	10	94.44
38	10	94.74	10	94.74
40	0	0	0	0.00

Table 2: This table summarizes the numbers of hits returned by blasting both the perfect and imperfect motifs initially found back against the genome. Only hits demonstrating a perfect alignment length and a percent identity score of 90% or greater are counted.

Total Number of Blast Hits (Perfect Palindromes)	*Total Number of Blast Hits (Imperfect Palindromes)*	*Number of Intragenic Hits (Perfect Palindromes)*	*Number of Intragenic Hits (Imperfect Palindromes)*	*Percentage of Intragenic Hits (Perfect Palindromes)*	*Percentage of Intragenic Hits (Imperfect Palindromes)*
1254	17411	538	10809	42.90%	62.08%
394	5552	162	4214	41.12%	75.90%
116	1693	40	1085	34.48%	64.09%
48	6075	12	4114	25.00%	67.72%
22	3022	3	1357	13.64%	44.90%
10	2186	0	467	0.00%	21.36%
8	1725	0	147	0.00%	8.52%
8	1348	0	40	0.00%	2.97%
8	1486	0	161	0.00%	10.83%
12	989	0	47	0.00%	4.75%
12	12	0	0	0.00%	0.00%
10	10	0	0	0.00%	0.00%
10	10	0	0	0.00%	0.00%

Table 3: This table outlines the number of hits returned by the blast search that are intragenic.

QueryLength	*Percent of Bases Between Genes (Perfect Palindromes)*	*Percent of Bases Between Genes (Imperfect Palindromes)*
14	56.91501481	36.55775905
16	58.45494924	23.14368696
18	64.75095785	35.11846164
20	73.125	30.89643025
22	85.95041322	54.42483867
24	100	78.18320465
26	100	91.20979614
28	100	96.87822976
30	100	88.91982308
32	100	95.02356924
34	100	100
36	100	100
38	100	100

Table 4: This table illustrates the relative number of bases involved in repeats that occur in intergenic regions of the genome.

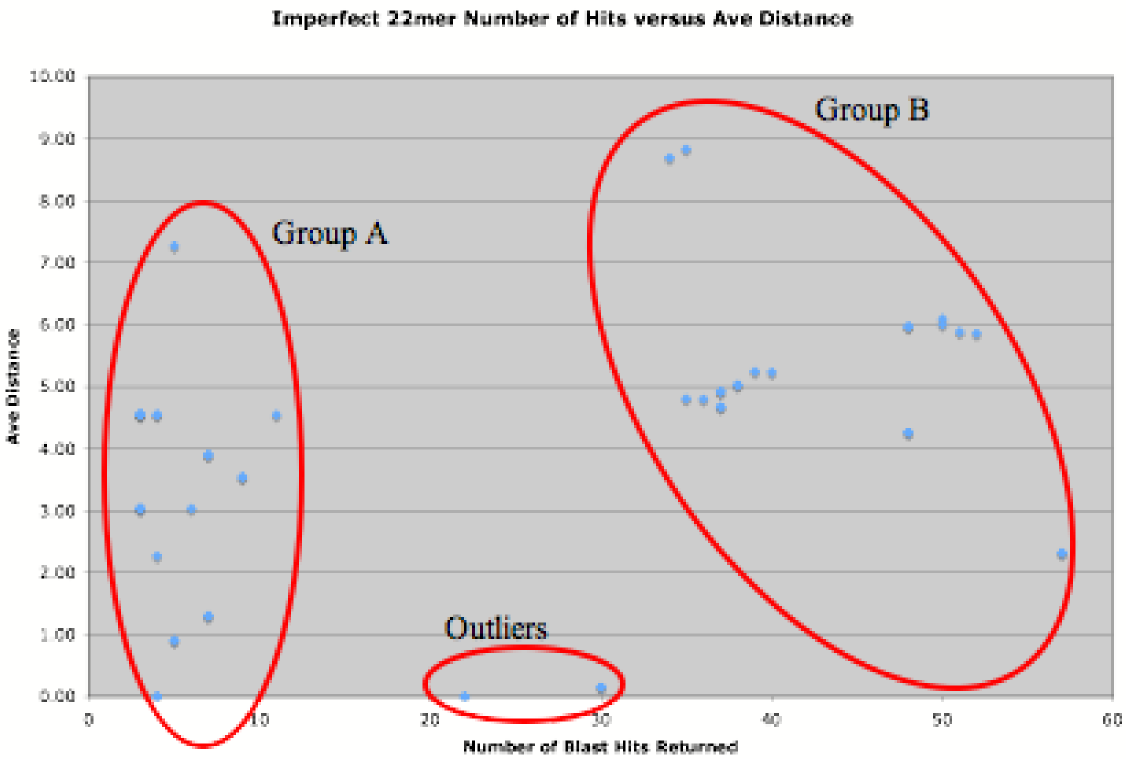


Figure 2: This chart shows the relationship between motif identity and number of motif copies of imperfect 22-base long palindromes in the genome. Each data point represents a specific motif. The “average distance” is calculated by averaging the percent identity scores determined by a Blast search for a certain motif and then subtracting this value from 100. Upon analyzing this data, two groups seem to emerge, as well as some outliers. Group A consists of 98 different motifs with low copy numbers, whereas Group B contains 24 different motifs with many copies in the genome. See the Supplementary Materials for the data values.

Length of Perfect Palindrome Query Sequences	*Number of Sequences*	*Difference Score*
14	70	0.68
16	23	0.70
18	2	0.67
20	1	1.00
22	1	1.00
24	1	1.00
26	1	1.00
28	2	0.79
30	1	1.00
32	1	1.00
34	1	1.00
36	1	1.00
38	1	1.00

Length of Imperfect Palindrome Query Sequences (Group A)	*Number of Sequences*	*Difference Score*	*Length of Imperfect Palindrome Query Sequences (Group B)*	*Number of Sequences*	*Difference Score*
14	3064	0.70	14	8	0.62
16	901	0.70	16	6	0.54
18	160	0.72	18	4	0.63
20	259	0.71	20	19	0.29
22	98	0.72	22	24	0.25
24	69	0.70	24	24	0.24
26	59	0.70	26	23	0.23
28	69	0.70	28	23	0.28
30	127	0.69	30	17	0.39
32	90	0.67	32	16	0.41
34	2	0.76	34	0	0.00
36	1	1.00	36	0	0.00
38	1	1.00	38	0	0.00

Table 5: These charts outline the number of different sequences of a motif of a given length that were blasted against the genome and returned between 2 and 25 hits (“Group A”) and more than 25 hits (“Group B”). The “Difference Score” is calculated by comparing a number of sequences base by base, adding up the number of mismatches between each pair of two sequences, and then dividing the total number of mismatches by the number of sequences. Thus, a lower Difference Score indicates less similarity among the sequences of a given length. The perfect palindrome sequences do not exhibit a significant trend of sequence similarity. The difference scores of the imperfect palindromes range from 0.67 to 0.76 in Group A, and they differ significantly in Group B, where they range from 0.23 to 0.62.

comprehensive data analysis, as well as limitations in evaluating the results. For this data set, the 90% identity threshold was applied, favoring sequences that diverged moderately from the initial query sequence. However, the 90% threshold can also skew Blast result numbers. For example, a certain sequence in the genome might not match the query sequence perfectly, but will be added to the data set on account of its percent identity score, which is greater than 90%. The data will be associated with this particular query sequence. However, the Blast program will also filter out the sequence on the reverse strand of the genome, which will most likely be within the 90% threshold of a different query sequence. This reverse sequence will then be associated with another query sequence. Therefore, one specific imperfect "sequence region" might be counted twice, once as the forward sequence and once as the reverse sequence (see Figure 3).

5' -- AAAAAAATTTT**G**TT -- 3'
3' -- TTTTTTTAAAA**C**AA -- 5'

Figure 3: The "palindromicity" of these two sequences is above the 90% threshold (13 of the 14 bases preserve the palindromic nature), but since the sequences have different patterns, this region in the genome will be found twice by Blast. Mismatches are highlighted in bold.

Also, as the length of a query sequence increases, the number of bases allowed to vary in a hit sequence increases, due to the threshold. This means, a 14-base query sequence yield a certain number of hits, but the same region in the genome might be hit again by longer query sequences (see Figure 4)..

5' -- AAAAAAATTTT**G**TT -- 3' 5' -- AGAAAAAAAATTTT**G**TT**G**T -- 3'
3' -- TTTTTTTAAAA**C**AA -- 5' 3' -- TCTTTTTTTTAAAA**C**AA**C**T -- 5'

Figure 4: The first 14-base pair sequence might be hit a certain number of times. The second, 20-base pair long sequence, is counted as a separate pattern, even though it contains the 14-base pair sequence. Mismatches are highlighted in bold.

3.5 Neighboring Genes

To illucidate possible functions of specific palindrome sequences in the genome, we extracted 2000 base pair sequences flanking the 14 base pair motif `ccaggcgcgcctgg`. Occurring seven times in the genome, this pattern is the most common perfect palindrome we found.

A Blast analysis revealed that four of these flanking sequences belong to genes known to be involved in the metabolism of amino acids in closely related *Pseudomonas* species (Table 6).

4 Conclusion

By utilizing a systematic analysis of the *Pseudomonas fluorescens* SBW25 species, we have investigated the existence of palindrome patterns in the genome. The statistical overabundance of these patterns suggests they may be important for the organisms survival and functionally significant.

Examining certain overrepresented patterns further highlights that their sequences are highly conserved throughout the genome. The fact that the organism is actively expending energy to preserve certain sequences leads us to believe that these patterns might represent common regulatory elements in the genome.

Finally, comparing flanking sequences of certain preserved patterns in the *Pseudomonas fluorescens* SBW25 genome to genes of fully annotated related species may reveal their function. Based on our observations we propose that one specific palindrome, `ccaggcgcgcctgg`, is involved in regulating the expression of genes that participate in the metabolism of amino acids.

Pseudomonas aeruginosa PAO1
Pseudomonas putida KT2440
Pseudomonas fluorescens Pf-5
Pseudomonas syringae pv. tomato str. DC3000
Pseudomonas syringae pv. phaseolicola 1448A
Pseudomonas syringae pv. syringae B728a
Pseudomonas fluorescens PfO-1
Pseudomonas entomophila L48
Pseudomonas aeruginosa UCBPP-PA14
Pseudomonas stutzeri A1501
Pseudomonas mendocina ymp
Pseudomonas putida F1
Pseudomonas aeruginosa PA7
Pseudomonas putida GB-1

Table 6: Fully sequenced *Pseudomonas* species included in the comparative genomics model.

5 Future Research

To infer functions of the remaining palindromic sequences we detected, we can perform an expanded comparative genomics analysis incorporating a set of fully annotated species closely related to *Pseudomonas fluorescens* SBW25. These predictions based on "in silico" analysis will promote direct traditional experiments in the laboratory using genetic manipulation of these sequences.

Acknowledgements

Dr. Allen Rodrigo (University of Auckland, New Zealand)
Dr. Paul Rainey (Massey University, New Zealand)
Dr. Philip Hatcher (University of New Hampshire, USA)
Dr. W. Kelley Thomas (University of New Hampshire, USA)

References

[1] Jansen, Ruud., van Embden, Jan.D.A., Gaastra, Wlm., Schouls, Leo.M. (2002) Identification of genes that are associated with repeats in prokaryotes. *Molecular Microbiology*, **43**, 1565-1575.

[2] van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H. (1998) Short-Sequence DNA Repeats in Prokaryotic Genomes. *Microbiology and Molecular Biology Reviews*, **62**, 275-293.

[3] Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L., Marqus, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida. Nucleic Acids Research*, **30**, 1826-1833.

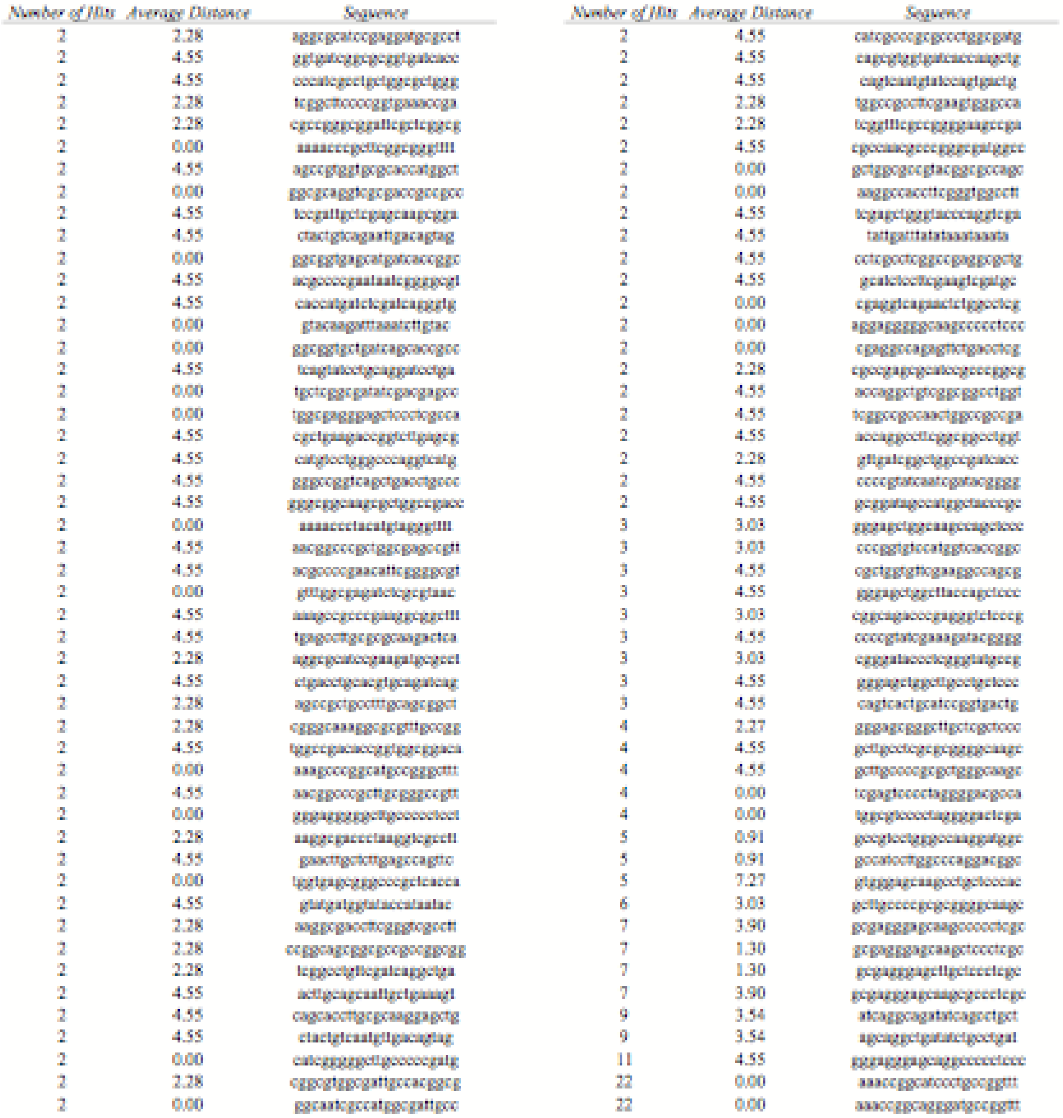
[4] Oxford English Dictionary.

[5] Pseudomonas fluorescens Project at the Wellcome Trust Sanger Institute.
`http://www.sanger.ac.uk/Projects/P_fluorescens/`

[6] The Perl Programming Language.
`http://www.perl.com/`

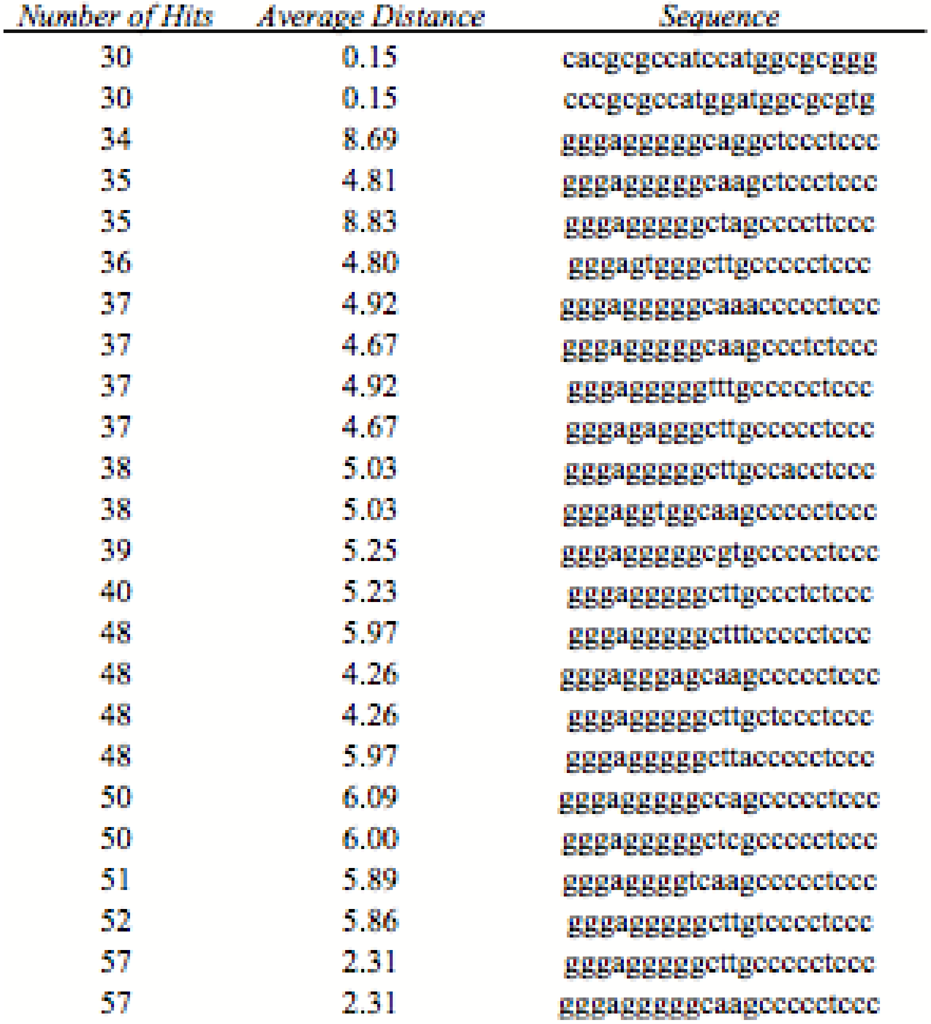
[7] Basic Local Alignment Search Tool (BLAST) at the National Center for Biotechnology Information.
`http://blast.ncbi.nlm.nih.gov/Blast.cgi`

Supplementary Materials



Number of Hits	Average Distance	Sequence	Number of Hits	Average Distance	Sequence
2	2.28	aggcgcatccgaggatgcgcct	2	4.55	catcgcccgcgccctggcgatg
2	4.55	ggtgatcggcgcggtgatcacc	2	4.55	cagcgtggtgatcaccaagctg
2	4.55	cccatcgcctgctggcgctggg	2	4.55	cagtcaatgtatccagtgactg
2	2.28	tcggcttccccggtgaaaccga	2	2.28	tggccgccttcgaagtggccca
2	2.28	cgccggcgcggattcgctcggcg	2	2.28	tcggtttcgccggggaagccga
2	0.00	aaaacccgcttcggcgggtttt	2	4.55	cgccaacgcccgggcgatggcc
2	4.55	agccgtggtgcgcaccatggct	2	0.00	gctggcgccgtacggcgccagc
2	0.00	ggcgcaggtcgcgaccgcgcc	2	0.00	aaggccaccttcgggtggcctt
2	4.55	tccgattgctcgagcaagcgga	2	4.55	tcgagctgggtacccaggtcga
2	4.55	ctactgtcagaattgacagtag	2	4.55	tattgatttatataaataaata
2	0.00	ggcggtgagcatgatcaccggc	2	4.55	cctcgcctcggccgaggcgctg
2	4.55	acgccccgaataatcggggcgt	2	4.55	gcatctccttcgaagtcgatgc
2	4.55	caccatgatctcgatcaggtg	2	0.00	cgaggtcagaactctggcctcg
2	0.00	gtacaagatttaaatcttgtac	2	0.00	aggagggggcaagccccctccc
2	0.00	ggcggtgctgatcagcaccgcc	2	0.00	cgaggccagagttctgacctcg
2	4.55	tcagtatcctgcaggatcctga	2	2.28	cgccgagcgcatccgcccggcg
2	0.00	tgctcggcgatatcgacgagcc	2	4.55	accaggctgtcggcggcctggt
2	0.00	tggcgagggagctccctcgcca	2	4.55	tcggccgccaactggccgccga
2	4.55	cgctgaagaccggtcttgagcg	2	4.55	accaggccttcggcggcctggt
2	4.55	catgtcctgggcccaggtcatg	2	2.28	gttgatcggctggccgatcacc
2	4.55	gggccggtcagctgaccggccc	2	4.55	ccccgtatcaatcgatacgggg
2	4.55	gggcggcaagcgctggccgacc	2	4.55	gcggatagccatggctacccgc
2	0.00	aaaaccctacatgtagggtttt	3	3.03	gggagctggcaagccagctccc
2	4.55	aacggcccgctggcgagccgtt	3	3.03	cccggtgtccatggtcaccggc
2	4.55	acgccccgaacattcggggcgt	3	4.55	cgctggtgttcgaaggccagcg
2	0.00	gttggcgagatctcgcgtaac	3	4.55	gggagctggcttaccagctccc
2	4.55	aaagccgcccgaaggcggcttt	3	3.03	cggcagacccgagggtctcccg
2	4.55	tgagccttgcgcgcaagactca	3	4.55	ccccgtatcgaaagatacgggg
2	2.28	aggcgcatccgaagatgcgcct	3	3.03	cgggatacccctcgggtatcccg
2	4.55	ctgacctgcacgtgcagatcag	3	4.55	gggagctggcttgcctgctccc
2	2.28	agccgctgcctttgcagcggct	3	4.55	cagtcactgcatccggtgactg
2	2.28	cgggcaaaggcgcgtttgccgg	4	2.27	gggagcgggcttgctcgctccc
2	4.55	tggccgacaccggtggcggaca	4	4.55	gcttgcctcgcgcggggcaagc
2	0.00	aaagcccggcatgccgggcttt	4	4.55	gcttgccccgcgctggcaagc
2	4.55	aacggcccgcttgcgggccgtt	4	0.00	tcgagtcccctaggggacgcca
2	0.00	gggagggggcttgcccccctcct	4	0.00	tggcgtcccctaggggactcga
2	2.28	aaggcgaccctaaggtcgcctt	5	0.91	gccgtcctgggccaaggatggc
2	4.55	gaacttgctcttgagccagttc	5	0.91	gccatccttggcccaggacggc
2	0.00	tggtgagcgggcccgctcacca	5	7.27	gtgggagcaagcctgctcccac
2	4.55	gtatgatggtataccatcatac	6	3.03	gcttgccccgcgcggggcaagc
2	2.28	aaggcgaccttcgggtcgcctt	7	3.90	gcgagggagcaagccccctcgc
2	2.28	ccggcagcggcgccgccggcgg	7	1.30	gcgagggagcaagctccctcgc
2	2.28	tcggcctgttcgatcaggctga	7	1.30	gcgagggagcttgctccctcgc
2	4.55	acttgcagcaattgctgaaagt	7	3.90	gcgagggagcaagcgccctcgc
2	4.55	cagcaccttgcgcaaggagctg	9	3.54	atcaggcagatatcagcctgct
2	4.55	ctactgtcaatgttgacagtag	9	3.54	agcaggctgatatctgcctgat
2	0.00	catcggggcttgccccgatg	11	4.55	gggagggagcaggccccctccc
2	2.28	cggcgtggcgattgccacggcg	22	0.00	aaaccggcatccctgccggttt
2	0.00	ggcaatcgccatggcgattgcc	22	0.00	aaaccggcagggatgccggttt

Table 7: Data set for imperfect palindrome motifs that are 22 bases in length in Group A (up to 25 hits in the genome).



Number of Hits	*Average Distance*	*Sequence*
30	0.15	cacgcgccatccatggcgcggg
30	0.15	cccgcgccatggatggcgcgtg
34	8.69	gggagggggcaggctccctccc
35	4.81	gggagggggcaagctccctccc
35	8.83	gggagggggctagccccttccc
36	4.80	gggagtgggcttgccccctccc
37	4.92	gggagggggcaaaccccctccc
37	4.67	gggagggggcaagccctctccc
37	4.92	gggagggggtttgccccctccc
37	4.67	gggagagggcttgccccctccc
38	5.03	gggagggggcttgccacctccc
38	5.03	gggaggtggcaagccccctccc
39	5.25	gggagggggcgtgccccctccc
40	5.23	gggagggggcttgccctctccc
48	5.97	gggagggggctttccccctccc
48	4.26	gggagggagcaagccccctccc
48	4.26	gggagggggcttgctccctccc
48	5.97	gggagggggcttaccccctccc
50	6.09	gggagggggccagccccctccc
50	6.00	gggagggggctcgccccctccc
51	5.89	gggaggggtcaagccccctccc
52	5.86	gggagggggcttgtcccctccc
57	2.31	gggagggggcttgccccctccc
57	2.31	gggagggggcaagccccctccc

Table 8: Data set for imperfect palindrome motifs that are 22 bases in length in Group B (more than 25 hits in the genome).