Analysis and Identification of Local DNA Structures

Ing. Michal Petrovič | Supervisor: prof. RNDr. Ing. Jiří Šťastný, CSc.

Department of Informatics, Mendel University in Brno

Motivation

The correct evaluation of local DNA structures is essential for understanding gene regulation and its implications in diseases. Z-DNA conformations and CpX Islands (extended CpG Islands) are critical for gene expression and epigenetic changes [2]. However, current existing tools lack the ability to effectively discover each of those structures, limiting research potential and speed.

This thesis aims to solve the problem by expanding DNA Analyser application to include tools for detecting Z-DNA structures and CpX Islands. Enhancing these competencies will provide researchers with better insights into the biological roles of the genome sequences and improve DNA analysis accuracy.

Conclusion

In the thesis, the functionality of the DNA Analyser web application was extended with new analytical tools specializing in Z-DNA conformations and CpX Islands. These improvements allow for extra efficient exploration of local DNA structures, facilitating studies in genomics and related fields. The primary objective was to develop, implement, and integrate new analyses that enhance the identification of specific sequences within the application, making it more available to the scientific community.

Theoretical research outlined the importance of Z-DNA and CpX islands in biomedical studies and explored available techniques for their identification. Based on those insights, new analytical tools have been developed and successfully integrated into the platform.

Testing confirmed the effectiveness of the new methods across diverse data sets, demonstrating their capability to identify wanted structures within the DNA sequences. Verification against the original models proved the correct calculations and identifications.

Future improvements should aim to integrate additional tools for identifying another type or adapting new algorithms to distinctive computational models. Furthermore, code refactoring and improvements to the user interface could provide an even more intuitive and efficient platform for DNA analysis.

Methods

The methods employed in this thesis involve the integration of two key algorithms to analyze Z-DNA structures and CpX islands.

CpX Islands

For the identification of CpX islands, the Takai and Jones' algorithm was adapted and extended [4]. This algorithm focuses on detecting regions with high dinucleotide frequencies (CpG, CpA, CpT and CpC) and uses a sliding window approach to identify islands with its full potential. The tool, CpX Hunter, allows for variable input parameters, enhancing its flexibility in experimental settings.

The algorithm begins by setting a sliding window of 200 nucleotides. Each window must meet certain criteria to be classified as an island. These include:

- Minimum C+X content of 50%.
- An observed-to-expected CpX ratio of at least 0.6.
- A minimal CpX dinucleotide percentage of 50% within the window.

Additionally, windows that are separated by less than 100 nucleotides are merged to form larger islands. These parameters help ensure that only regions of significant biological relevance are identified as CpX islands.

The formulas we use to determine if the window apply are these [3]:

$$C\_X = [Count(C) + Count(G)]/Length$$
$$ObsCpX = Count(CpX)/Length$$
$$ExpCpX = (C\_X/2)^2$$

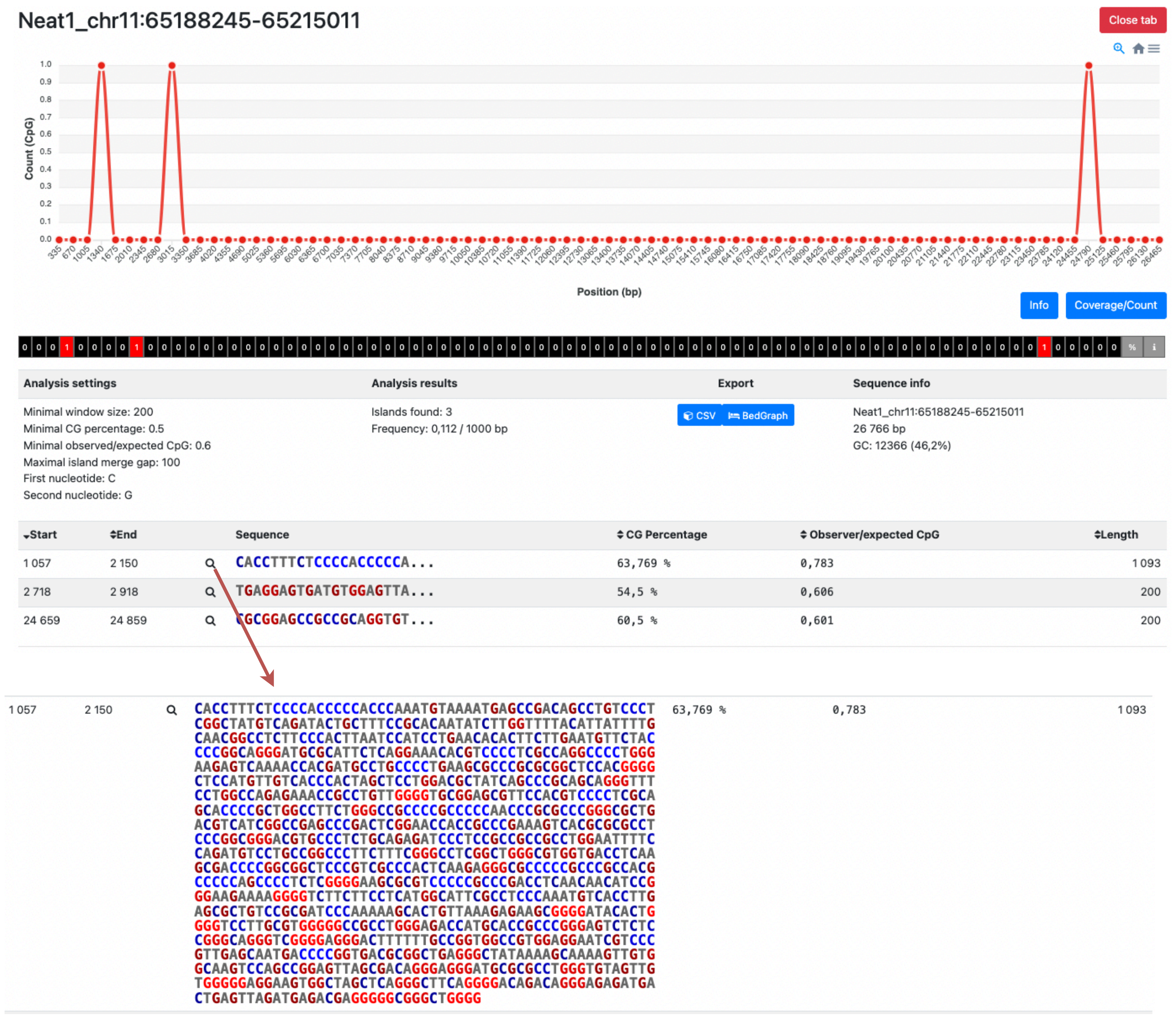
Z-DNA conformations

For the identification of Z-DNA conformations, the non-B-gfa tool [1] was customized to detect left-handed Z-DNA structures in genomic sequences, particularly in CG-rich alternating purine-pyrimidine regions.

Z-DNA Hunter calculates a score for each nucleotide based on its propensity to form Z-DNA. The scoring continues until a nucleotide that cannot form Z-DNA (e.g., AA, AG, TT, etc.) is encountered. If the window meets the minimum length and score criteria, it is marked as a potential Z-DNA region.

$$score\_perc = \frac{\text{score}}{(\text{length} - 1) \cdot \frac{\max(\text{score}_{GC}, \text{score}_{GT\_AC}, \text{score}_{AT})}{2}} \cdot 100$$

Results



Neat1_chr11:65188245-65215011

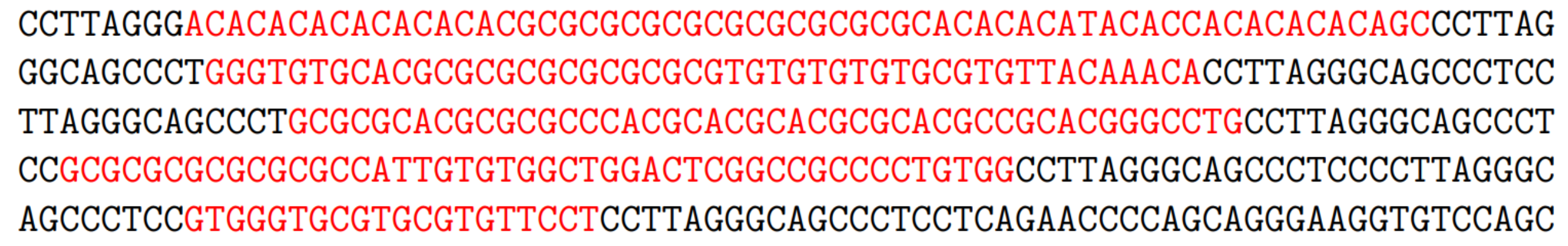
0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0

Analysis settings	Analysis results	Export	Sequence info
Minimal window size: 200 Minimal CG percentage: 0.5 Minimal observed/expected CpG: 0.6 Maximal island merge gap: 100 First nucleotide: C Second nucleotide: G	Islands found: 3 Frequency: 0,112 / 1000 bp	CSV BedGraph	Neat1_chr11:65188245-65215011 26 766 bp GC: 12366 (46,2%)

Start	End		Sequence	CG Percentage	Observer/expected CpG	Length
1 057	2 150	🔍	CACCTTTCTCCCCACCCCCA...	63,769 %	0,783	1 093
2 718	2 918	🔍	TGAGGAGTGATGTGGAGTTA...	54,5 %	0,606	200
24 659	24 859	🔍	CGCGGAGCCGCCGCAGGTGT...	60,5 %	0,601	200

1 057	2 150	🔍	CACCTTTCTCCCCACCCCCACCCAAATGTAAAATGAGCCGACAGCCTGTCCCT CGGCTATGTCAGATACTGCTTTCCGCACAATATCTTGGTTTTACATTATTTTG CAACGGCCTCTTCCCACTTAATCCATCCTGAACACACTTCTTGAATGTTCTAC CCCGGCAGGGATGCGCATTCTCAGGAAACACGTCCCCTCGCCAGGCCCCTGGG AAGAGTCAAAACCACGATGCCTGCCCCTGAAGCGCCCGCGCGGCTCCACGGGG CTCCATGTTGTCACCCACTAGCTCCTGGACGCTATCAGCCCGCAGCAGGGTTT CCTGGCCAGAGAAACCGCCTGTTGGGGTGCGGAGCGTTCCACGTCCCCTCGCA GCACCCCGCTGGCCTTCTGGGCCGCCCCGCCCCCAACCCGCGCCCGGGCGCTG ACGTCATCGGCCGAGCCCGACTCGGAACCACCGCCCGAAAGTCACGCGCGCCT CCCGGCGGGACGTGCCCTCTGCAGAGATCCCTCCGCCGCCGCCTGGAATTTTC CAGATGTCCTGCCGGCCCTTCTTTCGGGCCTCGGCTGGGCGTGGTGACCTCAA GCGACCCCGGCGGCTCCCGTCGCCCACTCAAGAGGGCGCCCCCGCCCGCCACG CCCCCAGCCCCTCTCGGGGAAGCGCGTCCCCCGCCCGACCTCAACAACATCCG GGAAGAAAAGGGGTCTTCTTCCTCATGGCATTCGCCTCCCAAATGTCACCTTG AGCGCTGTCCGCGATCCCAAAAAGCACTGTTAAAGAGAAGCGGGGATACACTG GGGTCCTTGCGTGGGGGCCGCCTGGGAGACCATGCACCGCCCGGGAGTCTCTC CGGGCAGGGTCGGGGAGGGACTTTTTTGCCGGTGGCCGTGGAGGAATCGTCCC GTTGAGCAATGACCCCGGTGACGCGGCTGAGGGCTATAAAAGCAAAAGTTGTG GCAAGTCCAGCCGGAGTTAGCGACAGGGAGGGATGCGCGCCTGGGTGTAGTTG TGGGGGAGGAAGTGGCTAGCTCAGGGCTTCAGGGGACAGACAGGGAGAGATGA CTGAGTTAGATGAGACGAGGGGGCGGGCTGGGG	63,769 %	0,783	1 093

Figure 1. Showcase of CpX Hunter for Neat chromosome 11.



```
CCTTAGGGACACACACACACACACGCGCGCGCGCGCGCGCACACACATACACCACACACACAGCCCTTAG
GGCAGCCCTGGGTGTGCACGCGCGCGCGCGCGTGTGTGTGTGCGTGTTACAAACACCTTAGGGCAGCCCTCC
TTAGGGCAGCCCTGCGCGCACGCGCGCCCACGCACGCACGCGCACGCCGCACGGGCCTGCCTTAGGGCAGCCCT
CCGCGCGCGCGCGCGCCATTGTGTGGCTGGACTCGGCCGCCCCTGTGGCCTTAGGGCAGCCCTCCCCTTAGGGC
AGCCCTCCGTGGGTGCGTGCGTGTTCCTCCTTAGGGCAGCCCTCCTCAGAACCCCAGCAGGGAAGGTGTCCAGC
```

Figure 2. Showcase of Z-DNA structure in human genome.

References

[1] R. Z. Cer, D. E. Donohue, U. Mudunuri, N. A. Temiz, M. Loss, N. J. Starner, G. N. Halusa, N. Volfovsky, M. Yi, B. Luke, A. Bacolla, J. R. Collins, and R. D. Stephens.
Non-b db v2.0: a database of predicted non-b dna-forming motifs and its associated tools.
Nucleic Acids Research, 41:D94–D100, 11 2012.

[2] P. S. Ho.
Thermogenomics: Thermodynamic-based approaches to genomic analyses of dna structure.
Methods, 47:159–167, 03 2009.

[3] S. Saxonov, P. Berg, and D. L. Brutlag.
A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters.
Proceedings of the National Academy of Sciences of the United States of America, 103:1412–1417, 01 2006.

[4] D. Takai and P. A. Jones.
Comprehensive analysis of cpg islands in human chromosomes 21 and 22.
Proceedings of the National Academy of Sciences, 99(6):3740–3745, 2002.

https://bioinformatics.ibp.cz

https://git.pef.mendelu.cz/bioinformatics