Unravel motifs in UTRs and introns

Andreas Jangmo[1,◕], Johan Lord[2,◕]

1,2 Royal Institute of Technology (KTH)/School of Computer Science and Communication (CSC), Stockholm, Sweden

◕These authors contributed equally to this work.

Abstract

A sequence logo is a graphical representation of conserved bases in a sequence of DNA or protein. It is similar to a bar graph with the bars being stacks of letters corresponding to the nucleotides or amino acids. The logo is created from a file of aligned sequences and the size of the letters correspond to the frequency of that base at the position in the sequence specified on the x-axis. The logo can be used to illustrate a specific motif or the presence of functional units or protein binding sites in DNA sequences. In this report we will study what the sequence logos are for the the regions before and after the translation start site and the first intron of each gene in the human genome. In doing so we will first extract the sequences for the regions of interest using Ensamble's Biomart. Using these extracted sequences we align them using python and create sequence logos using the python package Biopython. We end with a brief discussion of the interpretation of the resulting logos.

Introduction

In genetics, a motif is a pattern in nucleotide or amino acid sequences that have, or are thought to have, a biological significance. Motifs are important as they may determine a protein's secondary structure when present in exons. It may also indicate binding sites for a large variety of proteins such as enzymes or more direct RNA level processes such as ribosome binding. Some proteins bind with very high specificity such as Type II restriction enzymes [1]. These are part of the immune system in bacteria where they destroy viruses by splicing them up. Deviating from their binding sites could thus lead to catastrophic results. More common is that motifs vary more in composition such as the TATA box that indicates the binding site for RNA polymerase. It is apparently very rare to find a promoter that matches this sequence exactly [1]. A convenient tool when analyzing motifs is therefore to construct something like a histogram or a bar graph illustrating the frequency of each nucleotide at each position in a region of DNA from a large number of aligned sequences. This is done by creating something called a sequence logo [2]. A sequence logo is a graphical way of representing the variation of nucleotides or amino acids around a certain site that makes it easier to find candidate motifs. Each position relative to the site is assigned a score of information content, that is essentially a measure of the distance from a state of randomness that per definition has no information. The information content is measured in bits from complete random being 0 to perfect conservation represented by 2 bits as there are $2^2 = 4$ types of nucleotides. The actual letters representing the four nucleotides in our case are then scaled with

their information content at each position in the sequence thus resulting a diagram called a sequence logo.

The purpose of our work has been to create sequence logos for regions in the human genome. The regions we have focused on are the ones before and after translation start site and before and after the beginning and end of the first intron of those genes that has at least one intron. Creating sequence logos for these regions might hopefully reveal conserved patterns that indicate the presence of motifs.

Materials and Methods

Gathering gene data

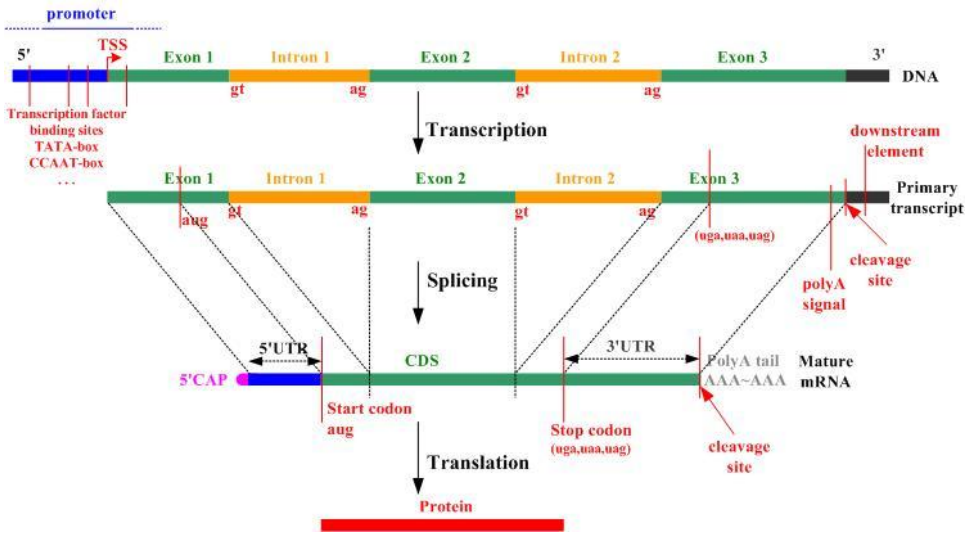

Figure 1. An illustration of a gene showing what happens to different regions through the processes of transcription, splicing and translation. [5]

Having decided on using as much of the human genome as possible we gathered the data using Ensamble's Biomart [3]. We have used the december 2013 Homo sapiens high coverage assembly GRCh38. At Biomart we choose the "Ensambl Genes 83" database with the dataset GRCh38.p5. We divided this work up into two parts: retrieving the sequences before and after the translation start site and retrieving the sequences from the beginning and end of the first intron.

Solving the first part was easy enough. From the dataset mentioned above (GRCh38.p5) we selected only coding sequences as these all start with the initiation codon. Further we added 15 nucleotides upstream to these sequences in our query to include an appropriate number of nucleotides before and after the start codon for our analysis.

In retrieving the introns this proved to be a bit more tedious as Biomart does not directly supply intron sequences. Studying the structure of a gene, seen in *Fig* 1, one could though come to some conclusions. A gene include both non coding parts such as UTR:s and introns, and coding parts which consist of parts of one or more exons. Exons include the 5' and 3' UTR:s and a number of exons are spliced out to assemble a transcript which are further spliced to become a coding sequence. Thus taking the difference of the complete genes and the exons will leave the introns. As we also want to include regions upstream the start of the introns as well as downstream the end of them we achieve this by using nucleotide coordinates. This method is also convenient for

other more obvious reasons as we are dealing with large amounts of data (all unspliced genes from the human genome accumulates to approx. 1.8 Gb of data). Picking out the sequences of interest by using coordinates limits the amount of data we would have to process and therefore greatly improves processing times.

Using this data we now continue by, more precisely filtering out the sequences to use as a base for the logo. We then need to align them and automate the logo creation. We implement all this using Python and more specifically the package Biopython [4].

Processing gene data

-HIS EXCELLENCY ANDREAS CONTINUES HERE BY EXPLAINING IN MORE DETAIL ABOUT OUR ALGORITHM FOR THIS-

Creating sequence logos

For sequence logo creation there are a few tools at our disposal. We choose to use Steven Brenner's WebLogo [6] for its simplicity and the possibility of accessing this tool via the function weblogo() in the Biopython class motifs. Weblogo is mainly used through their web application which we initially used for experimental purposes. It can also be used by downloading their source code, by using their python package or from a third party software such as Biopython. We choose to implement this using the above mentioned function weblogo() in Biopython. Weblogo uses a multiple sequence alignment as a basis for logo creation. Three file formats for these alignments are available: FASTA, ClustalW and Flat. Flat format means that the sequences are just listed on top of each other without sequence names or other header information. Since we had no use for header information at this point, Flat format was definitely most suited for our purposes.

Using the flat formatted alignment files created in the previous step a python program called createlogo.py goes through each line of the file and adds that sequence as a Biopython Seq object to a list. From this list a Biopython motif object is created and from this object the logo is created by connecting to the weblogo service via the function weblogo() in the motifs class. Thus one needs an internet connection to use this program. The output of createlogo.py is simply a PNG file displaying the logo.

Results and Discussion

Nulla mi mi, venenatis sed ipsum varius, Table **??** volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

LOREM and IPSUM Nunc blandit a tortor.

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

Sed ac quam id nisi malesuada congue.

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 1

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 2

3rd Level Heading. Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table **??** volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

LOREM and IPSUM Nunc blandit a tortor.

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Text.

Supporting Information

S1 Video

Bold the first sentence. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Text

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S2 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. D'haeseleer Patrik, What are DNA sequence motifs? Nature Biotechnology 24, 423 - 425 (2006)
2. Schneider, T.D. & Stephens, R.M. Sequence Logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100 (1990).
3. Ensembl release 83, December 2015 © WTSI / EMBL-EBI http://www.ensembl.org/index.html
4. Biopython 1.66, released on 21 October 2015. http://biopython.org/wiki/Main_Page

5. Prof B. Jayaram & Co-workers, Genome Tutorials. Supercomputing Facility for Bioinformatics & Computational Biology, IIT Dehli. As of 2016-01-03 avaible here: http://www.scfbio-iitd.res.in/research/genomics.html

6. Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004)

   http://weblogo.berkeley.edu/