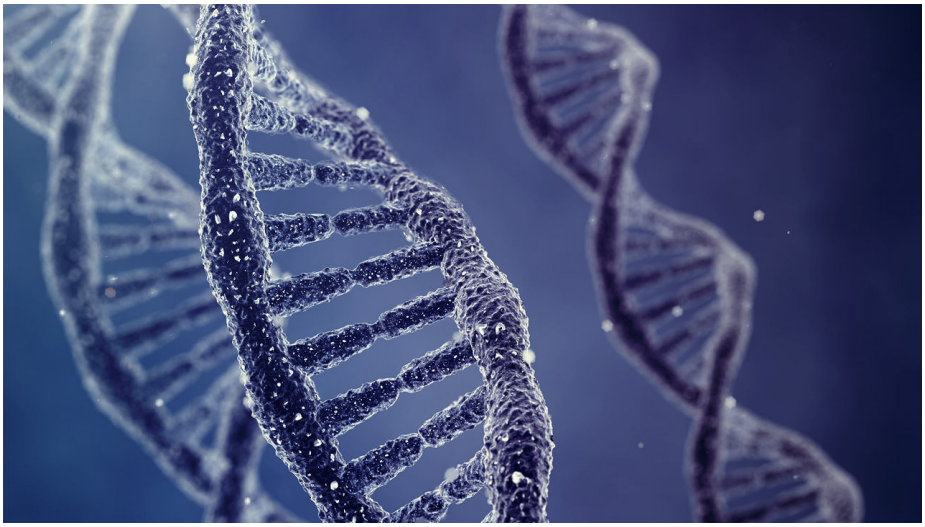
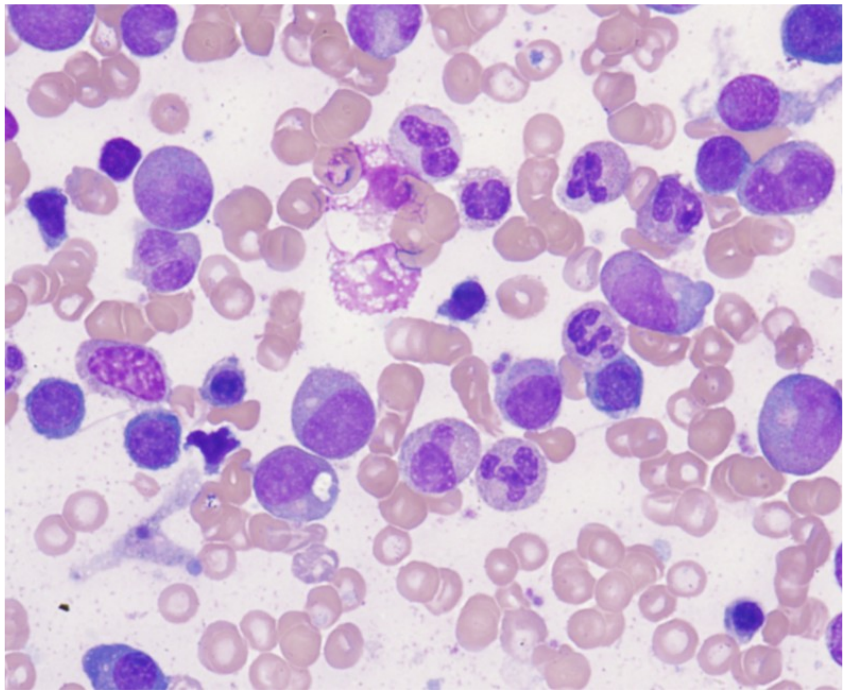


A Novel Algorithm for Identifying Sequence Motifs

By Advait Maybhate

Sir John A. Macdonald S.S.



Abstract

The project aims to develop a novel algorithm for finding sequence motifs. The goal of the project is to develop an algorithm that would be faster than existing algorithms without compromising accuracy. Motifs are short sequence patterns that represent fundamental units of biological function, such as recurring nitrogenous bases in DNA sequences. They encode protein, DNA, and RNA interactions, such as gene expression. Finding motifs allows biologists to predict the biological function of certain parts of the genome e.g. the NANOG transcription regulator motif has been linked to the pluripotency of embryonic stem cells. The new algorithm is centered around the concept of comparing a biological sequence with its randomly shuffled counterpart to identify significant motifs. The proposed algorithm proved to be very fast and quite accurate, as shown by the results obtained when testing it on previously gathered sequences, with known motifs. The algorithm was able to accurately identify prominent motifs in sequences containing the NANOG, STAT1, RUNX3 and C-jun motifs. When analyzing a PITX1 sequence, the “E-box” motif was shown to be occurring very frequently. Upon further research, it was found that the PITX1 motif has protein-protein interactions with the E-box motif. Interestingly, they did not co-occur together in the same sequence, rather the DNA strand itself was bent to allow these interactions. In conclusion, a novel algorithm for accurately and quickly finding motifs in DNA sequences was successfully created. This will have several applications, such as identifying mutations in the epsilon4 allele of the Apolipoprotein E gene (APOE) that can cause Alzheimer’s disease. Identifying which parts of the DNA/RNA sequence are causing such diseases can lead to the better diagnosis and treatment, potentially a cure.

Background Research

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)

DNA and RNA are nucleic acids, made from monomers (nucleotides) with three key components: a sugar (deoxyribose for DNA and ribose for RNA), a phosphate group, and a nitrogenous base (adenine, cytosine, guanine and thymine/uracil). DNA is the hereditary material in all forms of life, including humans. It carries genetic instructions used for growth, development and reproduction. RNA is a molecule essential to various functions of living organisms. Specifically, it is responsible for decoding, and regulating the expression of genes, including the creation of proteins.

In terms of representing sequences, DNA's nucleotide sequences are represented by: adenine (A), cytosine (C), guanine (G), and thymine (T).

Proteins

Proteins are made up of amino acids which are organic compounds containing amine and carboxyl functional groups. Proteins carry out various, essential functions within the human body from structural cell support to enzymatic activity. They are formed when DNA goes through the process of transcription to form a messenger ribonucleic acid (mRNA) and this mRNA is then translated to synthesize proteins.

Amino acid sequences are represented by either the name, three letter or one letter codes of the amino acids which they are composed from e.g. Lysine is represented as "Lys" or "K".

Motifs

Motifs are short sequences patterns that represent fundamental units of biological function, and they encode protein, DNA, and RNA interactions, as well as catalytic functions. These motifs can be found in both the coding and non-coding parts of the genome. Such motifs allow for the identification of genes that contribute to specific functions of an organism.

Descriptions of Motifs Analyzed

PITX1 (Paired Like Homeodomain 1) Motif

The PITX1 motif is part of the PITX protein family. This family is involved in organ development and left-right asymmetry. The sequence obtained for analysis was from the embryonic hind-limbs of a mouse, at an estimated 11.5 days of development. It is also thought to play a role in the development of anterior structures, specifically the brain and facial characteristics. PITX1 is also heavily involved in limb development. Defects in PITX1 can cause congenital clubfoot (foot abnormalities caused at birth) e.g. a twisted foot similar to the image shown below:

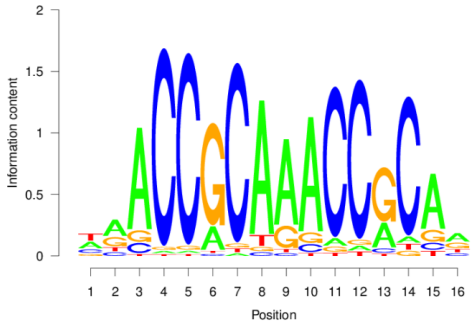
Generally, clubfoot is treatable if diagnosed early on. “GGATTA” is a very prominent binding motif for PITX1.

C-jun Motif

C-jun is a transcription factor that plays an important role in cell development, growth and differentiation. However, its binding sites and target genes are not clearly defined. C-jun interacts directly with specific target DNA sequences to regulate gene expression.

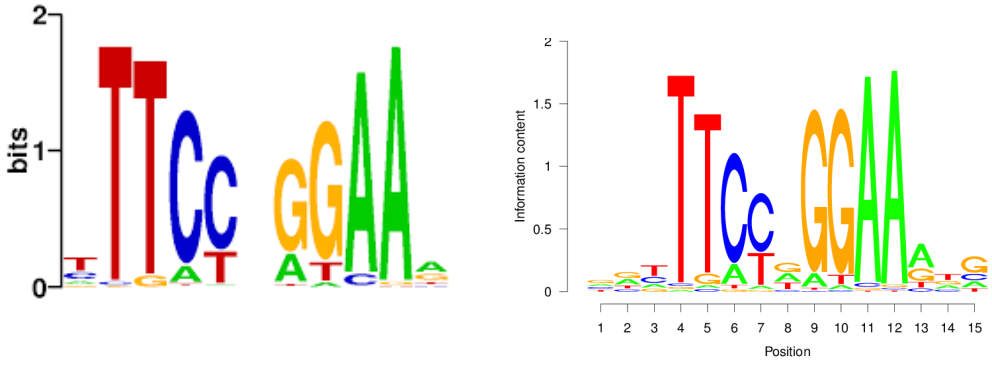
RUNX3 Motif

Runt-related transcription factor 3 is a protein that is coded by the RUNX3 gene. Specifically, those used in this project were taken from the GM12878 cell line which contains B-lymphocytes (white blood cells) infected with Epstein-Barr virus to achieve immortality (used to induce the expression of the RUNX3 transcription factor).



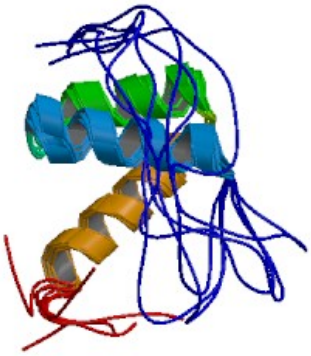
Stat1 Motif

The Stat1 transcription factor is essential to the cell cycle, especially during the cell transcription process in interphase, since it acts as a transcriptional activator in response to cytokines or growth factors.

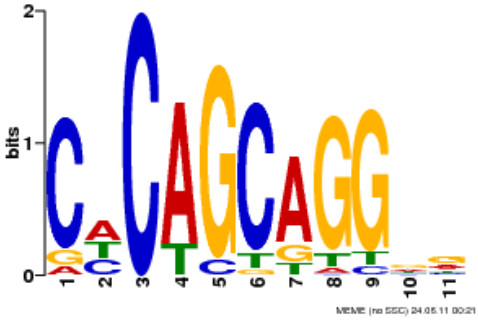


NANOG Motif

NANOG is a transcription regulator involved in embryonic stem cell proliferation (differentiation) and self-renewal, allowing for pluripotency. It is also involved in controlling inner cell mass.



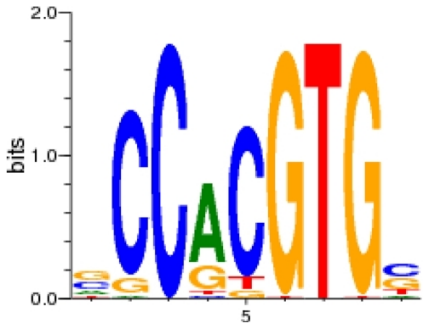
NANOG ribbon diagram



NANOG binding motif

E-box Motif

An E-box (enhancer box) is a DNA response element found in eukaryotes that acts as a protein-binding site. It is a basic helix-loop-helix (bHLH) in structure (characterized by two α-helices connected by a loop). It has been found to regulate gene expression in neurons, muscles, and other tissues. Its specific DNA sequence is: CANNTG (where N is any nucleotide). In this project, its prominent “CAGCTG” DNA sequence was analyzed (its other major variant is “CACGTG”).



ENCODE Project

The Encyclopedia of DNA Elements (ENCODE) is a public research project which aims to find all functional elements in the human genome. It was initially launched by the US National Human Genome Research Institute (NHGRI) in September 2003. It involves a collaboration of worldwide research groups and its data can be accessed through public databases. Many of the sequences used for this project’s testing were obtained from the ENCODE project e.g. sequences from embryonic stem cells containing the NANOG motif.

Purpose

The purpose of this project is to develop a novel algorithm which would attempt to solve the problem of finding DNA sequence motifs in an original way in order to increase the efficiency and accuracy of motif discovery. This will be achieved through the creation of a computer program that can analyze DNA sequences to extract significant motifs.

Design Criteria

A successful DNA motif discovery algorithm meets the following criteria:

- Accurately identify known motifs within an experimentally obtained set of DNA sequences
- Identify motifs faster than current algorithms e.g. the MEME (Multiple Expectation-maximization for Motif Elicitation) algorithm generally takes around 30 minutes to converge to specific motifs.

Procedure

The procedure for creating the DNA motif discovery algorithm is shown below:

1. Obtain large data sets containing DNA sequences. Many of the sequences used for analysis were obtained from the ENCODE (Encyclopedia of DNA Elements) project. These data sets were in the FASTA file format.
2. Parse the data contained within the FASTA files using Python (a computer programming language). This was achieved through the use of Biopython, a library of tools for biological computation.
3. Try different approaches to DNA motif discovery such as comparing a biological sequence with a random sequence to find unique motifs.
4. Refine the algorithm in order to produce more accurate results e.g. using the binomial test for statistical significance by finding the p-value.
5. Test the algorithm of a wide range of experimentally obtained sequences for its speed and accuracy.

Approaches to DNA motif discovery

- Analyze a DNA sequence for the frequency of k-mers (sequences of length k). Then, sort these results from highest frequencies to lowest frequencies. This approach was not very accurate when tested on several DNA sequences since it could not differentiate between “standard” biological motifs and specialized motifs.
- Compare a biological sequence with another sequence of the same size, however randomly generate the second sequence. The number of motifs found in the biological sequence will be compared against those found in the random sequence, to test for statistical significance. When tested on known

motifs within sequences, this approach had better success than the first, however it was not reliable. The frequencies of nitrogenous bases (A, C, G, or T) within the random sequence did not reflect those generally found in biological sequences
- Compare a biological sequence to its randomly shuffled counterpart. Such an approach ensures that the random sequence will have the same letter frequencies as the biological sequence, leading to a much more accurate analysis of the motifs.

Challenges

- Deciding whether to count one occurrence of each motif per sequence or multiple occurrences was a crucial question e.g. the motif “GAGCTG” may appear more than once in a sequence. After testing both strategies, it was found that only counting one occurrence resulted in more accurate motif discovery. It is theorized that this is due to compositional bias e.g. the string “AAAAAAAA” (length 8) counts as three 6-mers of the “AAAAAA” motif. This resulted in many low complexity motifs, such as “GGGGGG”, being unfairly ranked higher.
- Initially, the reverse complements of DNA sequences were not taken into account. When the reverse complements of sequences were combined in the algorithm, the prominence of desired motifs grew even larger.

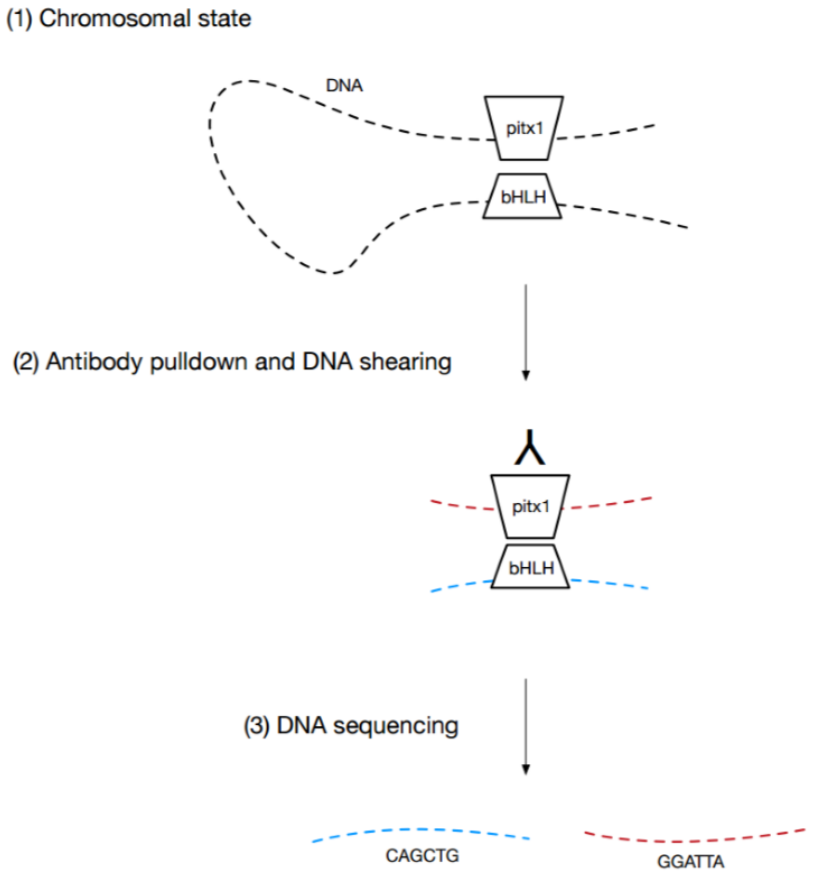
Results

PITX1 motif

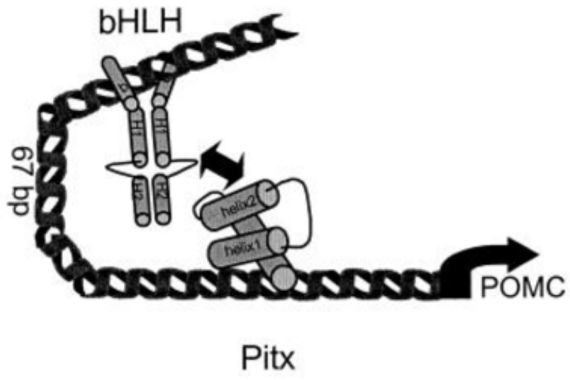
10 273 DNA sequences which were randomly sampled from the embryonic hind-limbs of a mouse, at an estimated 11.5 days of development. The motif which was desired was the “GGATTA” motif, however that was only ranked 18th. On the other hand, the top motif, which was “CAGCTG”, seemed to have many variations within the top 20 ranked motifs. When researched further, this was discovered to be the “E-box” motif (a bHLH — Basic Helix-Loop-Helix transcription factor). This was confirmed by the findings in previous scientific papers proving protein-protein interactions between these two motifs (Gino Poulin et al., 2000).[1]

Upon discovering the connection between the E-box and PITX1 motifs, an attempt was made to prove the co-occurrence of these two motifs, potentially finding a specific offset. However, when the data was analyzed to find co-occurring sequences, it was statistically insignificant as proven by the hypergeometric test. Therefore, a new model of protein-protein interactions was proposed:

[1] Poulin, Gino et al. “Specific Protein-Protein Interaction between Basic Helix-Loop-Helix Transcription Factors and Homeoproteins of the Pitx Family.” Molecular and Cellular Biology 20.13 (2000): 4826–4837. Print.



In this model, the DNA strand is bent to allow the PITX1 and bHLH motifs to interact. Interestingly, this model has also been proposed by a previous scientific paper (Gino Poulin et al., 2000):

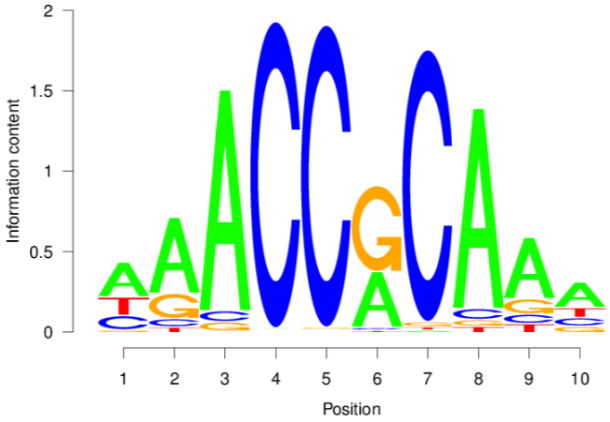


C-Jun Motif

The algorithm was able to identify “GAGTCA” which is very prominent in c-jun. The sample sequence is obtained from a human embryonic stem cell.

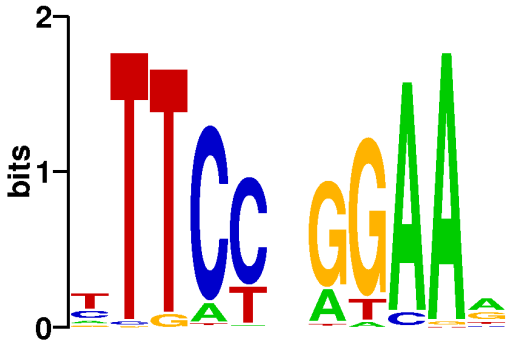
RUNX3 Motif

The desired motif was “ACCACA” which was ranked the highest by the algorithm, proving its accuracy. The sequences analyzed were taken from a human lymphocyte cell infected with the Epstein-Barr virus to induce the expression of RUNX3.



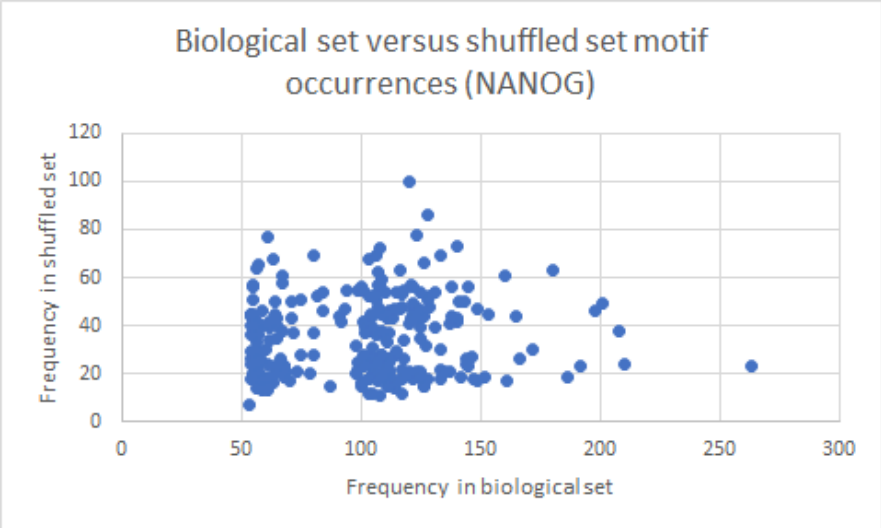
STAT1 Motif

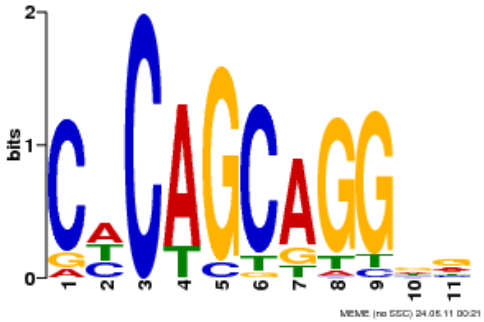
The STAT1 transcription factor is essential to the cell cycle. The algorithm was not confident in finding the motif and did not find the desired STAT1 motif, instead ranking “GGGCGG” as the most significant motif. This may be due to the fact that the STAT1 motif’s middle position is not certain (it can vary between the four nitrogenous bases) meaning that it would likely be better to find 4-mers for this motif.



NANOG Motif

The NANOG motif is involved in embryonic stem cell proliferation (differentiation) and self-renewal, making it essential to pluripotency of the cell. This sample sequence is taken from a human embryonic cell. The highest ranked motif by the algorithm’s analysis was “CAGCAG” which is very significant in the NANOG motif, proving the algorithm’s accuracy.





Conclusion

In conclusion, a newly-designed algorithm for DNA motif discovery was created successfully and was found to be relatively fast when compared to more complex algorithms, such as MEME, since analyzing data sets of over 10,000 sequences took around five minutes. This algorithm was coded in the Python language. It was able to accurately identify several motifs, in addition to predicting protein-protein interactions between the PITX1 and E-box motifs, which have been proven experimentally.

Applications

- Finding the biological function of different motifs
- Discovering new motifs
- Identifying which parts of the genome cause diseases
    - Can be used to repress disease-causing genes
- More than 98% of the genome does not code protein sequences. Its function remains mostly uncharacterized. Previously, it was thought that this was “junk” DNA with no function. However, recent studies have shown that this DNA is crucial in the process of transcriptional and translational regulation of protein-coding sequences.
- Non-coding DNA regulates DNA that codes for protein sequences by increasing or decreasing the expression of a target gene causing a specific feature/function to change, sometimes drastically.
- Certain parts of non-coding DNA/RNA can cause diseases/disorders in humans e.g. mutations in the epsilon4 allele of the Apolipoprotein E gene (APOE) occur in the non-coding regulatory region that facilitates transcription, can cause Alzheimer’s disease (a neurodegenerative disease). Identifying which parts of the DNA/RNA sequence are causing such diseases can lead to the better diagnosis and treatment, even potentially a cure.

Acknowledgements

I would like to thank Dr. Andrew Doxey for being my mentor and helping me understand biological terms and concepts.

Bibliography

Bailey, Timothy L. et al. “MEME: Discovering and Analyzing DNA and Protein Sequence Motifs.” Nucleic Acids Research 34.Web Server issue (2006): W369–W373. PMC. Web. 2 Apr. 2017.

"Basic helix-loop-helix." Wikipedia. Wikimedia Foundation, 23 Mar. 2017. Web. 02 Apr. 2017. <https://en.wikipedia.org/wiki/Basic_helix-loop-helix>.

"C-jun." Wikipedia. Wikimedia Foundation, n.d. Web. 01 Apr. 2017. <https://en.wikipedia.org/wiki/C-jun>.

D'haeseleer, Patrik. "What are DNA sequence motifs?" Nature News. Nature Publishing Group, 01 Apr. 2006. Web. 01 Apr. 2017. <http://www.nature.com/nbt/journal/v24/n4/full/nbt0406-423.html>.

"Encyclopedia of DNA Elements – ENCODE." ENCODE. N.p., n.d. Web. 01 Apr. 2017. <https://www.encodeproject.org/>.

Infante, C. R., Park, S., Mihala, A. G., Kingsley, D. M., & Menke, D. B. (2013). Pitx1 broadly associates with limb enhancers and is enriched on hindlimb cis-regulatory elements. Developmental biology, 374(1), 234-244.

Makolo, Angela. "A Comparative Analysis of Motif Discovery Algorithms." Computational Biology and Bioinformatics. Science Publishing Group, 20 Nov. 2015. Web. 02 Apr. 2017. <http://article.sciencepublishinggroup.com/html/10.11648.j.cbb.20160401.11.html>.

"NANOG." FactorBook RSS. N.p., n.d. Web. 25 Mar. 2017. <http://v1.factorbook.org/mediawiki/index.php/NANOG>.

"PITX1 Gene (Protein Coding)." GeneCards. N.p., n.d. Web. 02 Mar. 2017. <http://www.genecards.org/cgi-bin/carddisp.pl?gene=PITX1>.

"PITX1 (mouse)." Phosphosite. N.p., n.d. Web. 15 Mar. 2017. <http://www.phosphosite.org/proteinAction?id=2614419&showAllSites=true>.

Poulin, Gino et al. “Specific Protein-Protein Interaction between Basic Helix-Loop-Helix Transcription Factors and Homeoproteins of the Pitx Family.” Molecular and Cellular Biology 20.13 (2000): 4826–4837. Print.

"Structure of Nucleic Acids." SparkNotes. SparkNotes, n.d. Web. 02 Apr. 2017. <http://www.sparknotes.com/biology/molecular/structureofnucleicacids/section1.rhtml>.