NTNU

Kunnskap for en bedre verden

MOL3022 Bioinformatics

Identifying transcription factor binding sites with computational methods

Olaf Rosendahl

March 2023

1 Abstract

1.1 Background

All organisms store genetic information. Most organisms use DNA for this purpose [1]. Understanding genetics has been an essential part of the biology field, all the way from Gregor Mendel's experiments in the middle of the 19th century until today [2]. Important milestones have been the discovery of the DNA structure, published in 1953, and the Human Genome Project which was published in 2003. The Human Genome Project, which was started in 1990 and finished in 2003, mapped almost the entire human DNA [3]. This laid the foundation for DNA mapping and sequencing from different organisms. A lot of this information is stored in public databases [4]. The mapping of both human, and other organisms' DNA sequences, as well as the databases that store this information, is critical in making the use of computational methods to predict transcription factor binding sites possible.

Some of the most important work connected to understanding how DNA works and how genes are regulated is understanding transcription factors and their binding sites [4]. Understanding transcription factors and their binding sites are of vital importance to understanding gene regulation [5]. Using computational methods to identify transcription factors binding sites is both time- and cost-efficient [6].

1.2 Problem text

Develop a tool that can predict the most likely transcription factor binding sites in a DNA sequence using data on transcription factors from JASPAR and a given DNA sequence. The result should be visualized as a graph and list the positions of the likely binding sites in the sequence.

1.3 Results

This final result is a web-based tool for predicting likely transcription factor binding sites. Users can select a transcription profile from JASPAR and input one or more random DNA sequences. The tool does then show the results as a scatter plot and lists the likely sites.

1.4 Conclusion

The developed solution shows that a transcription factor binding sites predictor successfully can be implemented and give good results. When modifying the DNA sequence, the results clearly change, showing that the calculation works well. The solution can still be improved and especially more thorough testing should be considered for future work.

1.5 Keywords

DNA, Motif, Transcription factors, Transcription factor binding sites (TFBS), Position Weight Matrix, React, Typescript, JASPAR

2 Background

2.1 Genetic information

Genetics is stored in cells by nucleic acids called deoxyribonucleic acid, known as DNA [1]. Gregor Mendel conducted experiments in the 19th century that demonstrated how certain traits were inherited between generations [7]. The discovery of the DNA structure in 1953 by James Watson and Francis Crick [7] laid the foundation for the Human Genome Project, which began in 1990. In 2003, the project published a nearly complete sequence of the human genome [3], leading to a revolution in biological research and generating a vast number of genomics datasets stored in public databases [4]. Research then moved focus towards identifying functional elements [8].

2.2 Transcription factors and their binding sites

Transcription factors and their binding sites are two of the most important functional elements in the human genome. Transcription factors are proteins that bind to DNA-regulatory sequences, with their function being to modulate gene transcription rates, resulting in an increase or decrease in protein synthesis [5]. While many transcription factors are common to various cell types, some are cell-specific and these may determine the phenotypic characteristics of a cell [9].

Transcription factors bind to the promoter or distal regions in the DNA. Transcription factor binding sites are usually quite short, consisting of between five and fifteen base pairs [10]. By binding to regions, the transcription factors modulate gene expression [10]. A transcription factor has a preference for a specific set of DNA sequences, commonly referred to as binding motifs or transcription factor binding sites. Identifying these can be a challenging task, and various computational methods have been developed to facilitate the process.

2.3 Cataloging binding motifs

Most computational methods used for identifying transcription factor binding sites rely on a position weight matrix (PWM) which contains log-odds. These are used for computing the binding affinity score [10]. Cataloging the functional elements in the genome is an important part of understanding how cells differentiate [8]. By identifying transcription factors and their binding sites, researchers can understand how different cells respond to various stimuli and how specific cellular functions are regulated. This knowledge can have implications in other fields such as medicine, agriculture, and biotechnology [10].

3 Implementation

3.1 Technology

3.1.1 Typescript

As the task was about creating a tool, it was natural to create an interactive web-application. TypeScript is a strongly typed programming language that builds on JavaScript and compiles to JavaScript which means that the output can be executed in browsers [11]. For this project, Typescript was considered a better alternative than using standard JavaScript because of the advantages working with types gives, especially when writing algorithms with a lot of different variables in different structures.

3.1.2 React/Remix

React is a JavaScript library for creating user interfaces [12]. It allows developers to write HTML inside JavaScript together with an optional state. A state can for example store whether a button has been clicked or how many times it has been clicked. This makes it simple to create interactive interfaces. A big advantage of React is that it's very popular and a lot of functionality can be added to your website by simply importing packages.

Remix is a React-Framework for creating websites [13]. It offers easy usage of server-side rendering which means that the HTML is generated on the server before being sent to the user. Server-side loading can quicker since the hardware is more powerful, and thereby give a more smooth user experience. Generating webpages on the server also allows running functions and code which couldn't have been executed client-side because of technical or security limitations, for example, database connections. Another advantage is that it's easy to perform compute-intensive calculations at powerful servers before only sending the resulting visualization or result to the users.

3.2 TFBS calculation

When looking for likely TFBSs, one common strategy is to compare each possible motif in the given DNA sequence with transcription factors binding profiles. During this comparison, each sequence can be given a score that describes the probability of it being a TFBS. Here follows a step-by-step description of how this is done:

1. The position frequency matrix in the transcription factors binding profile from JASPAR is transformed to a position weight matrix (PWM) [14]. A PWM gives each nucleotide in each position a score that represents the log-likelihood of a nucleotide occurring at that specific position in the motif.
2. The given input DNA sequence is looped through, comparing each possible motif with the created PWM. The calculated score is found by adding the relevant values at each position in the PWM.
    - For example, if the nucleotide at position *5* in the possible motif is *T*. The score at *T[5]* in the PWM is used.

3. When each possible motif in the DNA sequence has been given a score, one can look at the scores to determine if it's likely that they are actual TFBS. The score indicates how different the possible motif is from a completely random sequence. If the score is 0, the probability of the sequence being a TFBS and a random site is the same. When the score is above 0, it is more likely to be a TFBS than a random site, and the other way around if the score is less than 0 [15].

4 Results

4.1 Web-application

The web application was designed with the goal of creating a simple and easy to understand user-interface and consists of three screens. The first screen (Figure 1) welcomes the users and describes the functionality of the application. It does also contain a list of the name and IDs of transcription factors binding profiles from JASPAR which the user can search and select from.

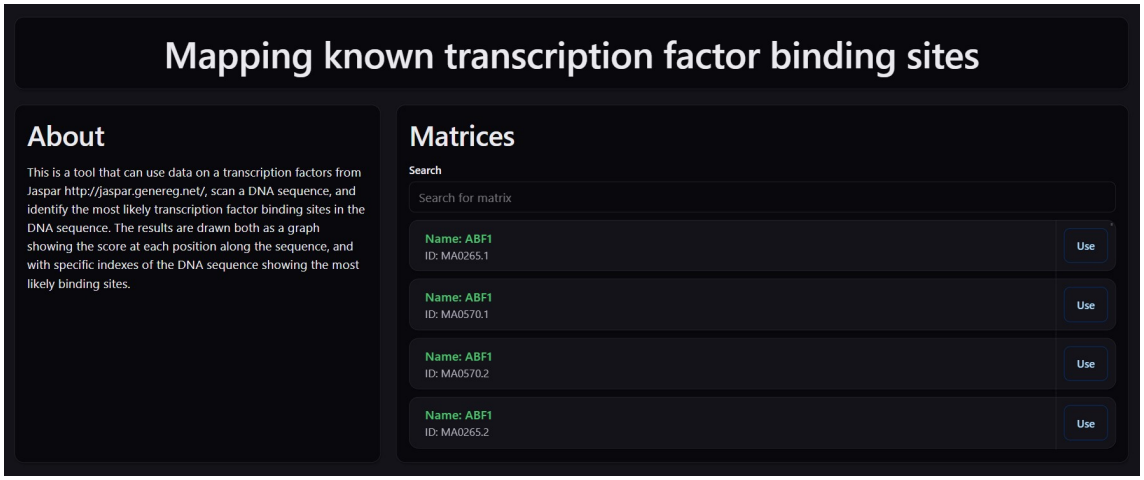

Figure 1: Screen 1 - About the website and selection of TF binding profile

The second screen (Figure 2) which is opened when the user selects on matrix shows more details about the selected matrix like its id, name, class, species, families, and type. In addition, the user can enter one or more sequences of DNA, separated with a comma for analysis with the chosen matrix.

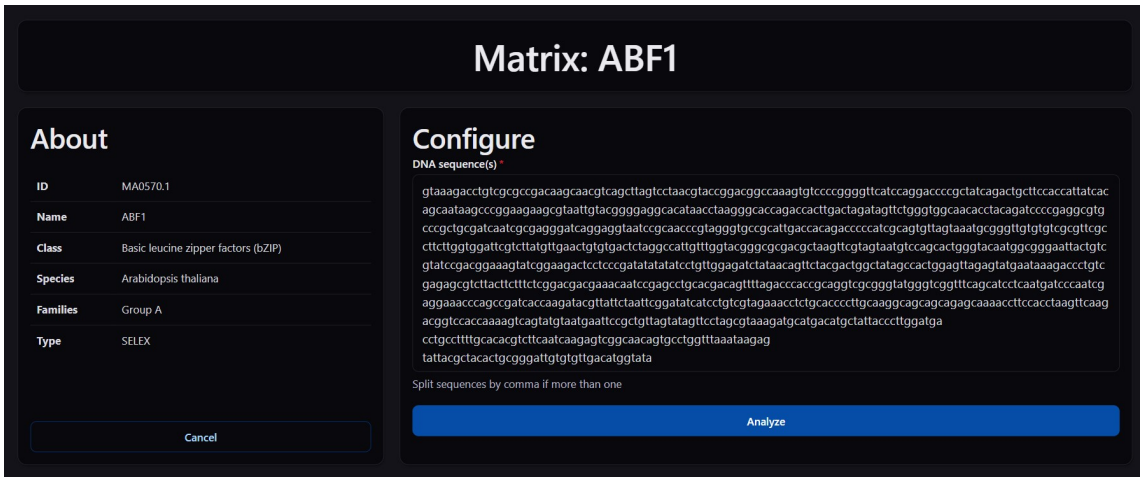

Figure 2: Screen 2 - View details about the selected TF binding profile and input DNA sequence(s) to analyze against it

When clicking "Analyze" on the previous screen, the selected matrix and the entered DNA sequences are sent to the server for calculations. When these are finished the user is redirected to the results screen (Figure 3). Here each of the entered DNA sequences is listed along with a scatter-plot that visualizes the scores which equal the likelihood of each position being a TFBS. There is also a list of the exact positions where the score is above zero.

Figure 3: Screen 3 - View the calculated results for each DNA sequence, visualized both by a scatter plot graph and with exact possible positions listed

4.2 Calculation

In order to test whether the implemented calculation provides valid results, a random DNA sequence was generated with Bioinformatics.org's Sequence Manipulation Suite [16]. The generated DNA sequence used was:

```
>random sequence 1 consisting of 100 bases.
ccccccccgtggcaccctctactattcatcacaatccgcgcacagatgac
ctcgcataaggggtcagtcgcttacggcggaggagtggttagttcgtagc
```

Then a random transcription factor binding profile was selected, `MA0016.1`. The sequence logo of MA0016.1, generated by JASPAR, shows the relative frequency of the corresponding bases at each position can be seen in Figure 4.

Figure 4: The sequence logo of MA0016.1, generated by JASPAR [17]

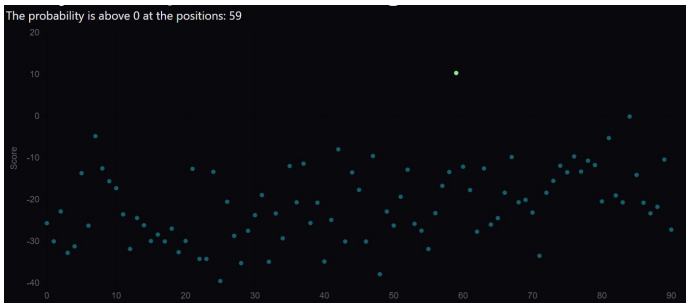

Figure 5: Result of calculation for the random DNA sequence with MA0016.1

By looking at the results in Figure 5 one can see that a possible motif was found at position `59` in the random DNA sequence. Figure 4 shows that MA0016.1 has a length of `10`, which means that by comparing the `10` nucleotides from position `59`, `ggggtcagtc`, it's possible to review whether the calculation actually has found a likely TFBS. When doing this comparison, we can see that the first 7 nucleotides, `ggggtca`, are equal in both. Then the rest of the positions, `gtc`, in MA0016.1 are evenly spread across all bases making it challenging to decide whether `gtc` is a good match, but it doesn't impact the result significantly.

When reversing the order of `t` and `c` at positions 5 and 6, and then run a new calculation on MA0016.1, one can see in Figure 6 that a match suddenly can't be found anymore. This corresponds with the fact that the sequence logo of MA0016.1 shows all matches have `t` at position 5.

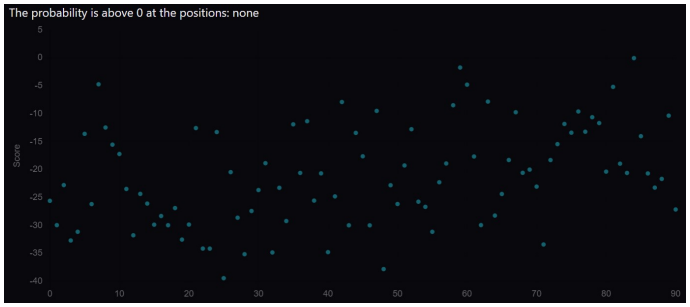

Figure 6: Result of calculation for the modified random DNA sequence with MA0016.1

5 Discussion

The created web application does the intended job well, making it easy for users to select a transcription factor binding profile, view information about it, input one or more DNA sequences, and view the calculated results for each input. Adding more textual information which informs users more in detail about what TFs, TFBS, and other keywords inside the application could have been useful depending on the target group. Displaying the actual matched part of inputted DNA sequence which gave a score above 0, instead of just the position, could also be useful and should be considered for future work.

Regarding the identification of likely TFBS in the DNA sequence, it seems that the results are fairly good. More testing could have been performed to conclude whether the algorithm scores the possible motifs correctly, but the simple testing shown in the Results does at least show that it manages to give completely separate scores when simply rearranging two nucleotides which is a good indication.

Overall in this project, the two students which have collaborated on the coding part of the project have done nearly all the coding together. I've focused especially on diving into how to implement an algorithm for finding likely TFBS, while the other student did some more on the user interface of the web application. In addition, much collaboration through pair programming has been done to ensure the correctness of the work done.

6 Conclusion

This paper has shown how a web application tool for predicting the most likely transcription factor binding sites in a DNA sequence has been built. The calculation implementation has been described in detail and the results from the testing show that it is capable of finding likely TFBS. These are also presented graphically with a scatter plot of scores and a list of where the likely TFBS is located in the input DNA sequence.

7 Availability and requirements

Instructions for running the developed application are included in `README.md` in the appended source code but a summary follows here:

1. Ensure Node.js version 16$\geq$ is installed
2. Install dependencies: `yarn` or `npm i`
3. Run application: `yarn dev` or `npm run dev`
4. Open `http://localhost:3000/` in a browser to use application

The application is also publicly available at `https://mol3022.vercel.app/` (may take some time to load due to the number of matrices to load from JASPAR).

References

[1] Walter Gilbert. "The road Not taken". In: *Nature* 320 (1986), pp. 485–486.

[2] Erik Dissen Lene Martinsen. "RNA". In: *Store Medisinske Leksikon* (2021).

[3] *The human genome project*. 2022. URL: https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project (visited on 02/13/2023).

[4] Francis S. Collins et al. "A vision for the future of genomics research". In: *Nature* 422 (2003), pp. 835–847.

[5] *Transcription Factor Definition*. 2019. URL: https://biologydictionary.net/transcription-factor/ (visited on 02/14/2023).

[6] Narayan Jayaram, Daniel Usvyat, and Andrew C. R. Martin. *Evaluating tools for transcription factor binding site prediction*. 2016. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1298-9#Sec1 (visited on 02/14/2023).

[7] Lene Martinsen. *Genetikk*. 2023. URL: https://snl.no/genetikk (visited on 02/14/2023).

[8] Martha L Bulyk. *Computational prediction of transcription-factor binding site locations*. 2003. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-5-1-201#citeas (visited on 02/13/2023).

[9] Ian M. Adcock and Gaetano Caramori. "Asthma and COPD". In: *Asthma and COPD* (2009). URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/transcription-factor (visited on 02/13/2023).

[10] Valentina Boeva. *Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells*. 2016. URL: https://www.frontiersin.org/articles/10.3389/fgene.2016.00024/full/ (visited on 02/14/2023).

[11] *TypeScript: JavaScript with syntax for types*. 2022. URL: https://www.typescriptlang.org/ (visited on 02/20/2023).

[12] *React - A Javascript library for building user interfaces*. 2022. URL: https://reactjs.org/ (visited on 02/20/2023).

[13] *Remix - Build Better Websites*. 2022. URL: https://remix.run/ (visited on 02/20/2023).

[14] Benjamin Jean-Marie Tremblay. "Introduction to sequence motifs". In: (Oct. 2021). URL: https://www.bioconductor.org/packages/devel/bioc/vignettes/universalmotif/inst/doc/IntroductionToSequenceMotifs.pdf.

[15] Roderic Guigo, IMIM/UPF/CRG. *An Introduction to Position Specific Scoring Matrices*. 2022. URL: https://bioinformaticaupf.crg.eu/T12/MakeProfile.html (visited on 02/20/2023).

[16] *Random DNA Sequence*. 2022. URL: https://www.bioinformatics.org/sms2/random_dna.html? (visited on 02/21/2023).

[17] *Matrix profile: usp - MA0016.1 - JASPAR*. 2022. URL: https://jaspar.genereg.net/matrix/MA0016.1/ (visited on 02/21/2023).