Tuberculosis Project

Hunter Casillas 10/18/19

Abstract

Mycobacterium tuberculosis is a species of pathogenic bacteria which is the causative agent of tuberculosis. It can persist in a latent state in humans for many years before causing disease and has been linked to hypoxia. With advances in modern computing power, we can use algorithmic programs to analyze the upstream sequences for all of the mycobacterium tuberculosis bacteria genes that are upregulated in hypoxia. Doing so makes it possible to find the motif that corresponds to the transcription factor regulating these genes. Through these programs, I was able to find the most probable motif for the binding site of the mycobacterium tuberculosis. Then I was able to find a motif in the 8-amino acid peptides that is most likely to be toxic to this bacteria.

Methods

In order to calculate the most probable motif for mycobacterium tuberculosis I used the 250 bp upstream region of each gene from the supplied fasta file, which my biologist colleague told me to look at. My colleague also told me that the motif is probably about 20 bp long (20 became my k value in my algorithm) which gave me a good starting point. In order to find the most probable motif I used the Gibbs Sampler function I had previously implemented. The function uses a random number generator to select a profile-randomly generated k-mer at each step. If the die rolls the number i, then we define the Profile-randomly generated k-mer as the ith k-mer.

After finding the most probable motif in mycobacterium tuberculosis, I was able to use it to find the motif in the peptide that is most likely to be toxic to our tuberculosis bacteria. I used the pandas python data analysis library to read the peptidesWithDNA.csv file given to us. I filtered my data from the file to only include the DNA sequences of the peptides classified as toxic. I then used the same Gibbs Sampler function described above to find the most probable motif for the antibiotic.

Results – Mycobacterium Tuberculosis and Antibiotic

The most probable motifs I found for mycobacterium tuberculosis and for the antibiotic were CCACCGCCGCCATCGCGCCG and TTTCTTTTTCTTTTTCTTTT, respectively. In order to calculate the mycobacterium tuberculosis motif I ran the Gibbs Sampler function two thousand times and kept track of the resulting best motif and score after each iteration. I checked to see if the new results were better than my previous ones and updated my answer as necessary. There were many other motifs found for mycobacterium tuberculosis through my program and they are listed below:

GCACGGCGCCCAGCCATTCG, GCGGGGCGCGGAGTTCGCCG, CCCGCCCGGACAGCCATGCG, CCCTGTCCGCGTCCGTGTCG, CGATGGACACCGTGGTCGCG, CCAGAGCCGTCGGCAGCCAG, CGGTCGTGGTCACTGCGGAG, GCCCGGCCGCCATCGGCCCG, CCATCGCCGCGGTCAAGCCG,

CGCCGGCAGCGTGCGTGACG, GCGCCGTGGCGGTGACAACG, CGGCGGCGGCGATACCCCAG, CGACCGCAGAAAGGGGGCAG, GCACCGGCGCTTTGGCGAGG, CCACACCGGCGACCGTTAGG, CCACCGCCGAGATCGCGGCG, CGATGGTGGGCGGCGGTCCG, GCCGCATCGGTGGCACCCCA, CGATCCTCGGCATCGGGCCG, CCCTGGCCACGATGGGCTGG, CCATCGTGGCCAGGGCTAGG, CCCCCGTTACCGTCGTGCGG, CCGAGCACGCCATTGTGCAG, CGACCCTAGTGTTCGCTACG, CCACCTCGGTCATCGACCCC, GCAGCCCGGCCAGCACGCCG, CGACAGCGAAGAGATCACCG, CCACGGCTGGCGATGTGCGG, CCAACGCCGCGAAAAACCGA, GCACCGAGGGTGTTTCGCGG, CGACAGCCGCGGTGGCAGAG, CCACCGCGGCTGTCGACGCG, CCATCGCCGGCGGCAGTGCG, CCGCCCCAGCGAAGGAGACG, CGCGCGGCGCCGGCTTGTCG, CCGTGTCGGAAATTAGGGCG.

I used a consensus function to find the most probable motif from my other possible motifs. This function used motif counting and built my final result of CCACCGCCGCCATCGCGCCG. I believe my method justifies my resulting motif because it was the best found after 2000 iterations. When it comes to the antibiotic I used basically the same procedure I did for mycobacterium tuberculosis motif with some slight modifications. I installed and used the pandas python data analysis library to read the peptidesWithDNA.csv file given to us. This was a quick and easy way of accessing the peptide data. I filtered my data from the file to only include the DNA sequences of the peptides classified as toxic. I used random sampling from the toxic peptides to run my Gibbs Sampler function on. I also used the consensus function to construct my final answer of TTTCTTTTTCTTTTTCTTTT. When comparing this result to the results of my classmates I found that almost everyone had very similar motifs. The majority of the results were constructed of T's and C's and were also 20-mers. Some results were very close in hamming distance to mine and on group even got the exact same result as I did, which shows that my answer is accurate. Additionally I know my Gibbs Sampler algorithm works because I've run it on test sequences from Rosalind and I get the correct results every time.

Discussion

When comparing my results to the linked articles I found that there are more powerful algorithms that can be used. The algorithm I implemented uses randomization and therefore isn't the most efficient one out there. In the article "Genome-Wide De Novo Prediction of Cis-Regulatory Binding Sites in Mycobacterium tuberculosis H37Rv" they state that they found a very reliable algorithm for identifying true binding sites for a specific TF. They used what is known as the eGLECLUBS (GLobal Ensemble CLUsters of Binding Sites) algorithm. In the eGLECLUBS algorithm first you extract upstream inter-operonic sequences of operons in each COOR to form a set of sequences. For each sequence set, you predict a total of T motifs using multiple motif-finding tools. Then you construct two motif similarity graphs G1 and G2 using a low and a high motif similarity score cut-offs α and β, respectively, and cut G2 into dense subgraphs using MCL. Next, you construct G3 and cluster it into dense subgraphs using MCL. Finally you induce subgraphs s1, s2, ..., sn of G1 using the clusters from G3, and identify quasi-cliques in each of these induced subgraphs. This algorithm is complex, but is very reliable and efficient. It really is amazing what the power of modern computing has done for bioinformatics. As computing power continues to increase each year, we will hopefully be able to eradicate illnesses and other genetic problems from the human population and improve the overall quality of life worldwide.