Finding regulatory modules using a linear HMM and the Viterbi algorithm

December 17, 2003
CSE527 Computational Biology
Project Report

Raphael Hoffmann

Introduction

To fully understand the function of genes in higher eukaryotes, one has to know the complex regulatory mechanisms that control gene expression. It is well-known that finding transcription factor binding sites can be a key to "crack" these mechanisms. Therefore, many techniques and algorithms that tackle this task have been developed and are available. However, their capabilities are limited: One reason is that short transcription factor motifs tend to occur frequently outside promoter regions, resulting in a large number of false positives. Another reason is that usually multiple transcription factors act in concert. Their motifs have to be aligned within a certain distance and often ordering. These cis-regulatory modules (CRM's) are typically a few hundred base pairs long.

Finding regulatory modules

A common approach to finding regulatory module, is to consider statistical significance of transcription factor binding motif clusters. MCAST, which is part of MetaMEME[1], uses exactly this idea. As input, it requires a set of motifs, each with its position-specific probability matrix (PSPM) and its occurences in a DNA sequence. This information can be obtained by running MEME[2], a motif discovery tool, on a DNA sequence. MCAST then builds a Hidden Markov Model (HMM) according to this data. The HMM represents a statistical model for a module, and contains an intra-module spacer state, an inter-module spacer state and each motif as a state. The Viterbi algorithm then uses the HMM to find statistically significant sites.

This general approach is not new, however, the authors have improved it in a number of ways: The HMM can be created using a linear topology, a star topology or a complete topology. The latter being more flexible, but requiring many parameters. Another main improvement solves a problem that all HMM's have: Since transitions are Markov, gap lengths are usually distributed geometrically which is not true for real data. The authors created a sophisticated scoring function that among other things allows arbitrary distributions for spacer lengths. Additionally, they tried to learn the transition probabilities of the HMM using the EM algorithm. However, they mentioned that this is not yet feasible due to the limited amount of reliable data that is available.

[1] MetaMEME - Motif-based hidden Markov modeling of biological sequences is available from http://metameme.sdsc.edu/
[2] Multiple EM for Motif Elicitation (MEME) can be downloaded from http://meme.sdsc.edu

Architecture

My application is object-oriented and based on JAVA. It is a simplified version of the approach that has been described. It only considers linear HMMs and uses the standard log-odds scoring function. The following figure shows an overview.

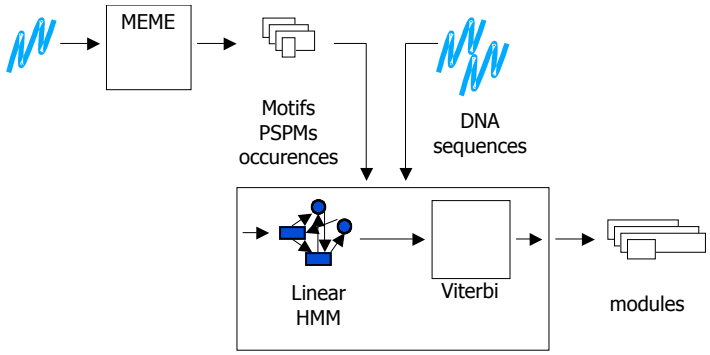


Figure 1. *Overview of the architecture*

MEME delivers a set of motifs with PSPMs and their occurences. Based on that information, a Linear HMM is built and the Viterbi algorithm applied. It delivers the most probable path through the model and its probability.

Testing on Simulated Data

Since results on real data depend on a large number of factors, it is recommendable to first test the performance of the algorithm on simulated data. I developed a small program that uses a given HMM and generates a random sequence according to background noise and the HMM. The actual states that served for generating the sequence are being stored in a log file for later comparison with the results of the module finding algorithm.
Of course, the probability of detecting the module correctly in the simulated data depends on the PSPM's of the motifs, the background distribution and especially the length of the motifs and the number of motifs.
In the following example I used two very short Transcription Factor binding motifs and connected them sequentially through a Linear HMM. The used motifs are explained later, when the same HMM will be applied to real data. The entire sequence is 16,000 bp long.
I used a sliding window and calculated the log-odds score in each position. The results are shown in the following figure

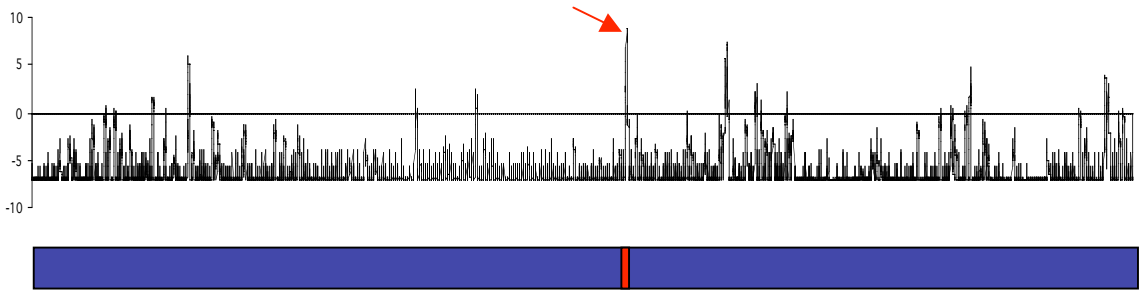


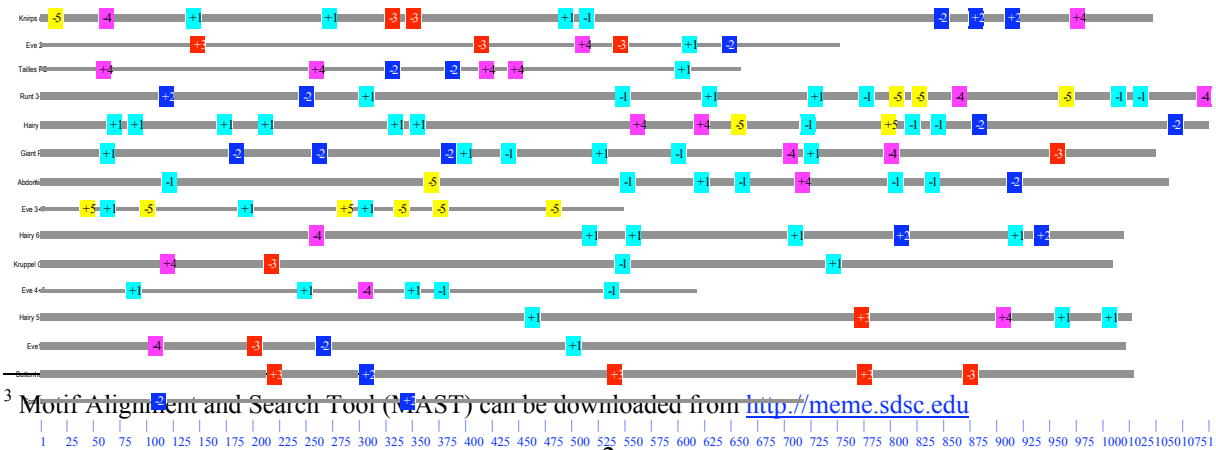
Figure 2. *The above diagram shows the log-odds scores. Below, blue indicates that the data was generated from the background model. Red indicates, it was generated by the HMM.*

The motif was found, i.e. it had the highest score.

Real Data

Testing the developed algorithm on real data turned out to be much harder than on simulated data. The reasons are various. There is only very limited well known data about regulatory modules available. Most research work focuses on one of the following two datasets: There are about 20 Drosophila developmental genes of which some CRM sequences are known. Those sequences can be obtained from [4]. Another dataset is from human, however I focused on the Drosophila dataset.
Some research in this field is obviously inspired by a simple and successful experiment [3]. About 1-Mb of DNA around the even-skipped gene was scanned using a 700-bp window. Sites that contained at least 13 predicted Transcription Factor binding sites (Bcd, Cad, Hb, Kr and Kni) correlated with known CRM sites unexpectedly well. In this experiment the relative ordering and distances between the binding sites was entirely ignored.
I first tried to extract the binding site motifs and their relative ordering from the sequences that contained known CRMs (available at [4]) using MEME. However,
MEME did not precisely detect the known motifs (Bcd, Cad, Hb, Kr and Kni). I used "meme –dna –revcomp –mod tcm –minw 8 –maxw 11 –nmotifs 5". However, even helping MEME by adding sample motif sequences, restricting the CRM sites and tweaking the parameters did not lead to success. Probably some motifs are not so clear from a statistical perspective.

I then used PSPM's for the five motifs which were available from [5], and explicitly aligned the motifs and the known CRM's using MAST[3]. The alignments are shown in the following figure:



[3] Motif Alignment and Search Tool (MAST) can be downloaded from http://meme.sdsc.edu

There is hardly any pattern in the alignment of the different motifs. Therefore, it makes no sense to apply the presented algorithm with a large HMM to find regulatory modules.
However, there is obviously a large amount of Hb-binding sites, and they often appear in pairs, e.g. in the first CRM. An obvious question is: Can this fact be used to identify the regulatory modules? I developed a simple HMM (following figure) and ran the algorithm on larger sequences.

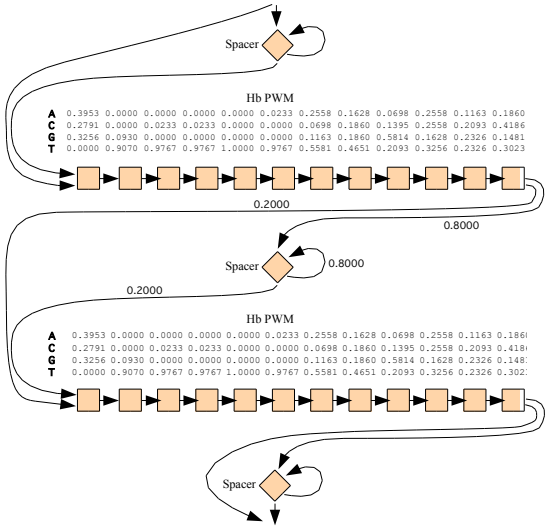


Figure 4. *The linear HMM used. It contains the Hb-motif twice.*

I used a sliding window and calculated the log-odds score in each position around the Hairy gene. The results are shown in the following figure.

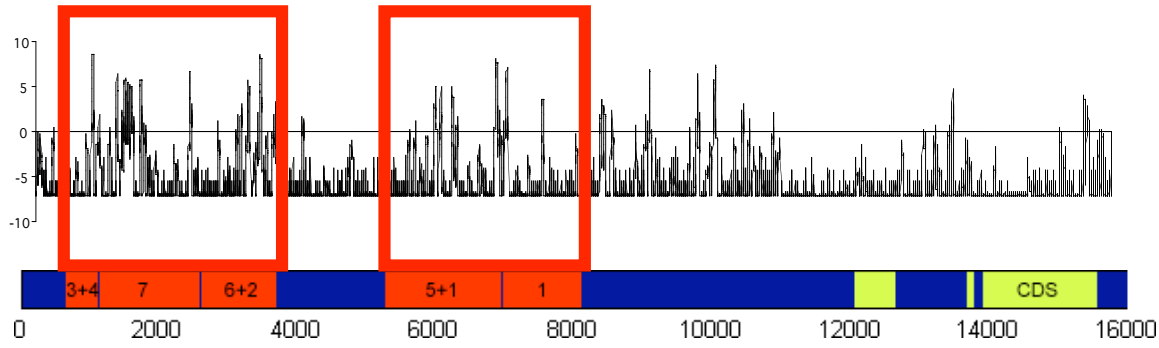


Figure 5. *Log-odds scores. Below, blue indicates that these regions are believed to be noncoding or that their function was not yet discovered. Red regions represent known regulatory modules. The green regions are exons.*

The results are not very clear. However, the scores in the exons (green) tend to be lower than average, while the scores (at least in the in the first 3 CRM's) are above average. There seems to be a region around 10,000 bp that has rather high scores, but has not been previously identified as a CRM.

Discussion

Interesting observations can be made by comparing the results of the same HMM on simulated data and on real data. On average, the log-odds scores on the real data were significantly higher than those on the simulated data, even in noncoding regions.

However, in general it is not possible to reliably detect CRM's using the simple HMM described above. Reasons:

- I only regarded motifs on one strand. However, the MAST experiments already showed that some patterns appear on the opposite strand. To include both strands, the HMM topology has to be extended.
- Also visible in the MAST results is that there is hardly any significant sequence of motifs that appears in multiple CRM's. Therefore, the Linear HMM approach fails. A Complete HMM or Star HMM could improve results.
- Since I used a classical HMM, the length of spaces between motifs is distributed geometrically. This model is definetly wrong, since short gaps will always be preferred over longer ones. A solution would be to associate individual arbitrary distributions to gap lengths.
- The biological processes are probably not fully understood and therefore no model is absolutely accurate. The large number of available papers that deal with finding the modules on the same small dataset (described above) shows that the task is not trivial.
- I simply used the log-odds score as a scoring function. [1] and others have defined sophisticated scoring functions that span hundreds of lines of code and that include many tweakable parameters.

Future Work

Timothy Bailey and William Noble [1] proposed to use the EM-algorithm to learn the parameters of the Hidden Markov Model, i.e. the transition probabilities. However, since few transcription factor binding sites are exactly known, the learning approach is infeasible at the time. As more and more data becomes available, the learning performance could certainly improve. I am convinced that learning the transition probabilities could become crucial, because the parameter values (e.g. transition probabilities) are now based on MEME output. This however could be often unreliable. Furthermore, MEME itself uses thresholds for the alignment and doesn't even output occurences with weaker agreements. This data could still be valuable for defining the right HMM. Applying a training method could solve these problems.

References

1. Bailey, T. L., Noble, W. S. (2003) Searching for statistically significant modules. *Bioinformatics*, vol 19, ii16-ii25.

2. Bailey, T. L. (2003) Details of MCAST Statistics. Technical Report IMB-TR0001. Institute for Molecular Bioscience, University of Queensland.

3. Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S., Levine, M., Rubin, G. M., Eisen, M. B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *PNAS* vol. 99 no. 2, 757-762.

4. Lifanov A., Makeev V., Nazina A. and Papatsenko D.(2003) Interactive collection of cis-regulatory modules from Drosophila. http://homepages.nyu.edu/~dap5/PCL/appendix2.htm

5. Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E. D. (2002) Computational Detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*. http://www.biomedcentral.com/content/pdf/1471-2105-3-30.pdf