Comparative Motif Finding
CS 374 – Lecture 23
Mayukh Bhaowal

Reference papers


Introduction

Recent studies have shown that most functional elements in the genome are non-coding and comprise regulatory elements that define when and where genes are expressed. Understanding the regulation of gene expression is a major challenge of current genomics, and the ability to identify regulatory motifs is an important step towards solving this issue. Computational methods are proving to be of great value to tackle this issue.

Background

1. What is a motif?

Motifs are simply nucleotide sequences and they tend to have biological significance. Regulatory motifs are nothing but DNA fragments.

2. Motif logos:

In the area of motif finding, motif logos, such as the one shown next, are important ways of representation.
The x-axis in the figure above represents the positions in the motif: 0 is the transcriptional start site, negative coordinates are upstream of it, and positive coordinates are downstream of it. The y-axis represents the bits of information that ultimately translate into the probability of finding a given letter in that position. For example, in position 8, the probability of finding a T is half the probability of finding a G, which in turn is half the probability of finding an A.

Why is it difficult to find motifs? First of all, motifs are generally short fragments, so it’s like finding a needle in a haystack problem. Second, they are degenerate. Generally, two or more bases can be found in a given location. Lastly, motifs can be unpredictable for several reasons: they can occur in either of the two strands of DNA and at varying distances from the transcriptional start. These are the main reasons why regulatory motif finding is an active area of research and the issue has not been solved.

3. **Promoter**: a DNA sequence that enables a gene to be transcribed. The promoter is recognized by the enzyme RNA polymerase, which then initiates transcription.

| Human       | CTGCCT------AAGTAGCTAGACGCTCCCGTGCG--CCC GGGGGGG--TAG |
| Mouse       | CGCCGC------CTGCATTATTAC--------------------------- |
| Rat         | CTGCTC------ATGCATAATTAC--------------------------- |
| Dog         | CTGCTTTCACAGTG GGGCAGACGTTCCCGCCGC GC CCAAGCAGGCCC |

4. **3’-UTR (three primer untranslated region)**: is a region of the messenger RNA that does not code for protein and therefore is not translated.

**First paper**: “*Systematic discovery of regulatory motifs in human promoters and 3’UTRs by comparison of several mammals*”

The first paper proposes to find motifs specifically in promoters and 3’-UTR of human genes.

The authors compare sequences of different mammals and try to find conserved sequences. Thus, their method is called **comparative motif finding**. For the purpose of the experiment, they consider four mammals: human, mouse, rat and dog.
Conservation properties

This is an evolutionary tree, which represents the sequence conservation in promoters and 3'-UTR respectively, in the four species. The figure on each branch shows the number of substitutions per site. Using these trees, we can understand the evolutionary history: the less sequence substitutions, the closer the two species are.

Methods

The authors resort to RefSeq, which is a well known curated sequence database and choose 17,700 genes from it. The promoters are defined as the 4kb regions upstream of the transcriptional start, while the 3'-UTR is derived from the annotation of the reference mRNA. In order to control for possible conservation due to chance, they select 123 Mb of intronic sequences.

Motif Conservation Score (MCS)

A motif is said to be conserved when an exact match is found in all four species.

The MCS is defined as:

\[
\text{Observed conservation} - \text{random conservation} \over \text{Standard deviation}
\]

where conservation is the number of conserved occurrences divided by the number of total occurrences.
For example, Err-α [TGACCTTG] is a highly conserved motif, as of the 434 times it occurs in human promoter regions, 162 of them are conserved across all the four species. Thus we can calculate a conservation rate of 37%, and compare it to that of a random 8-mer motif, which shows only 6.8%. In the following multiple alignment, one instance of the Err-α motif is highlighted. Asterisks denote those bases that are conserved across the four species.


In sum, the higher the MCS, the higher the conservation. The metric they use in the paper is that a motif is conserved if it has a MCS greater than 6.

Results

For the promoter region, the authors identified 174 highly conserved motifs, according to the above criteria, of which 69 match—at different degrees—previously known motifs. The remaining 105 motifs represent potential functional motifs (which have not been validated experimentally).

There are two parameters that can be used prove or infer functionality: tissue specificity and positional bias. The authors show that there is a high correlation between a high MCS and a high tissue enrichment score (which denotes tissue specificity). Likewise, for most motifs with high MCS, there is a very high bias to be located within
100 bases of the transcriptional start site (TSS). However, some highly conserved motifs like Err-α do not have a positional bias.

For the 3’-UTR regions, the authors found 105 motifs. Motifs in this region have not been studied before, so it’s not possible to compare the results to experimental data. However, they detect some interesting properties for these 3’-UTR motifs: first, contrary to motifs found in promoters, they have a very clear strand specificity, as can be seen in the following graph, where blue dots represent motifs from promoters and red dots motifs from 3’-UTRs. The axes show the conservation score in both strands. The second property is that most of the 3’-UTR motifs are 8-mers or overlap conserved 8-mers, in the case of shorter motifs. It’s important to remark that the authors are not only looking for 8-mers, but they take that as a start or seed, but the resulting motif can be longer or shorter.

As a digression, **microRNAs** are single stranded RNAs, transcribed from DNA but not translated into protein. Many mature miRNA start with a uracil (U) followed by a 7-base “seed” complementary to a site in the 3’-UTR of target mRNAs. Thus many miRNA motifs are 8-mers. The authors think that the fact that their 3’-UTR motifs are enriched for 8-mers is not a coincidence and they infer that they might be binding sites for miRNAs. Please note that they do not validate this assertion in the paper.
Second paper: “Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting”

The second paper deals with phylogenetic footprinting as a means of finding motifs. The key aspect that defines the method of phylogenetic footprinting for motif finding is that it considers only orthologous regulatory regions across several species. Paralogous and orthologous sequences are both homologous sequences. As represented in the figure below, orthologous sequences belong to different species and are derived from speciation from a common ancestor. They retain their functional properties in the two species. In contrast, paralogous sequences arise by duplication in one species and the copies do not retain the same functional properties.

A main concept that should be taken into account is that coding sequences evolving at a slower rate than non-coding sequences cause selective pressure. Mutations occurring in a coding sequence can possibly alter the whole function of the encoded protein, while mutations in a non-coding sequence, like a regulatory element, may only change the expression profile of a gene (levels or tissue specificity). Thus, we can study the non-coding sequences and this gives us better information to find a regulatory element than the coding sequence, because in this case the whole function will be altered.
In sum, phylogenetic footprinting is based on the study of orthologous non-coding DNA from species that are related (as represented by a phylogenetic tree) and leads to the identification of well conserved sequences that are putative regulatory elements.

There are differences between this approach and that of the previous paper: most importantly, in this approach we have a phylogenetic tree, so we know the relationship between the species. The second difference is that the scope of the authors of the previous paper is to find as many motifs as possible, whereas in this case the authors try to find one motif in many species.

In the algorithm for phylogenetic footprinting, which follows a dynamic programming (DP) approach, we are given a phylogenetic tree $T$ and a set of orthologous sequences at the leaves of that tree. Specifically, the problem is to find each set $S$ of $k$-mers, one $k$-mer from each leaf, such that the parsimony score of $S$ in $T$ is at most $d$.

Here we have an example:

```
AGTCGTACGTGAC... (Human)
AGTAGACGTGCCG... (Chimp)
ACGTGAGATACGT... (Rabbit)
GAACGGGAGTACGT... (Mouse)
TCGTGACCGGTGAT... (Rat)
```

This is the solution for $k=4$, with a parsimony score of 1:

```
AGTCGT
ACGT
ACGT
ACGG

AGTCGTACGTGAC...
AGTAGACGTGCCG...
ACGTGAGATACGT...
GAACGGGAGTACGT...
TCGTGACCGGTGAT...
```

There is one sequence for each species, connected following the tree of the species. At every leaf, potential mutations are placed. In this particular case, one leaf has one mutation and the other has none. We then ask what is the minimum number
of mutations that can explain the tree or, in other words, reconstruct all the internal nodes. That total number of mutations is the parsimony score. In this case is 1, because there is only one change. The aim is to minimize that score, over all possible collection of words, which is an exponential problem.

\[ W_u[s] = \text{best parsimony score for subtree rooted at node } u, \]
if \( u \) is labeled with string \( s \).

For each node (●), \( u \) is the label and \( s \) is the string. If \( u \) is a leaf node (see detail below), and the sequence ACGT is a substring of the entire sequence \( s \), we give a score = 0. Meanwhile, if a given string doesn’t appear, as in the case of ACGT, the score would be +∞. This is the case when \( u \) represents a leaf node.

What happens when \( u \) is not a leaf node? In that case, we sum it up across all its children, according to the recurrence formula below. In simple terms, the k-mer score at a node is the sum of its children’s best parsimony scores for that k-mer.
\[ W_u[s] = \sum_{v: \text{children of } u} \min ( W_v[t] + h(s, t) ) \]

For instance, if we consider the node at the bottom right, which has two children (see detail below): if we have the sequence ACGG at the parent node, which is the same one that appears in the children (or in other words is a substring of the child string), the score is 0; while if we place ACGT, one change has to occur in each of the children, so the total score is 2.

In that same node, the parsimony score of the string GAAC would be 2, because the bottom leaf would need 2 changes and the top one needs zero.

For a particular size \( k \), this algorithm computes a full table of \( 4^k \) scores for all nodes, starting from the leaves to all internal nodes, which is an exponential problem.

Below, the running time for this algorithm is represented. The total complexity is given by the expression on the right, where \( n \) is the number of species, \( k \) is the motif length and \( l \) is the average length of all sequences.
Results:

The authors analyze their results for several datasets. We will consider the three most important ones.

The first is the **Metallothionein Gene Family**, which comprises a large number of promoter sequences and regulatory elements, which occur within 300 bp of the start codon of the genes. Applying phylogenetic footprinting to the region comprising 590 bp of sequence upstream of the start codon and looking for conserved elements of lengths 7,8,9,10 (k values), the authors identified 12 motifs, of which 4 have been experimentally validated. The figure below shows the distribution of the 12 motifs in the promoters of different species. The DP algorithm, for any given node, apart from calculating the parsimony score, has a penalty for losing the motif in the whole subtree.

![Diagram showing distribution of motifs in different species](image)

The second dataset is the **Insulin Gene Family**. The authors take into consideration two gene copies for three different species (two rodents and a pig), and look for motifs with 0 mutations of length (k) 8 and motifs with 1 mutation of lengths 9 and 10. They identify 4 conserved motifs. This shows one of the drawbacks of their algorithm, as several binding sites are missed because they contain very few mutations.
For motifs with zero mutations, there are more efficient methods, since this algorithm is exponential. Even with 1 to 3 mutations, hashing would be a better approach.

The last dataset is the **C-myc family**, of which they analyze several species from diverse animal phyla (fishes, birds, mammals and frogs). Out of the 9 predicted motifs, 4 are known binding sites and most were located within 120 bp of the promoter region.

**Drawbacks of the method:**

- Some binding sites do not have significant matches to most other species (for instance, thyroid hormone receptors have matches in species like mouse and rat, but not in others).
- Some binding sites show good conservation rate in sequences shorter than footprinter looked at. Some are really short.
- Deletions/Insertions. The authors mention that their algorithm can take this into account, but they chose not to because that leads to a high false positive rate.
- Failure to meet statistical significance.
- Some transcription factors (TF) bind as dimers where the binding site may consist of 2 conserved regions of a given length, separated by a few variable nucleotides. Those instances would be missed. They have to be contiguous.