

Discovery of Conserved Regulatory Elements in Gene Promoter Regions

Meelis Kull
Department of Computer Science
University of Tartu
Meelis.Kull@ut.ee

Seminar "Bioinformatics of Gene Regulation" (MTAT.03.192)

December 20, 2005

Abstract

Regulatory elements are short substrings in the DNA which play a key role in gene regulatory processes. Finding these is essential for understanding the cell and the causes for various diseases. This short paper lists some possible ways to discover regulatory elements in the genome and describes the method by Elemento and Tavazoie (Elemento & Tavazoie 2005) in more detail.

1 Introduction

DNA is full of different type of information. The first type discovered were the genes, encoding the information for the production of proteins but also different types of RNA. The second level of information, in some sense, is achieved by the regulatory elements which affect the way gene information is interpreted.

Regulatory elements are short substrings in the DNA which are able to increase or decrease (or in the extremes, enable or disable) the activity of genes depending on whether there is a protein bound to the element or not. The effect of the element can sometimes appear only after transcription into mRNA. In that

case, some protein or some other RNA can affect the mRNA to be spliced alternatively or to be cleaved. Sometimes more than one regulatory elements on the DNA are known to work in combination to produce the effect.

2 Discovering Regulatory Elements

A lot of regulatory elements have been identified in the genomes of different organisms, although probably many more are to be found. The obvious way for discovery of such elements are the experiments indicating the act of binding. However, both, in vitro and in vivo experiments may not detect binding because we might not have some condition which is obligatory for binding. This suggests trying in silico methods for discovering the regulatory elements.

In the case of in silico methods one has to define the specifics of the object being searched for. In other words, we have to describe how regulatory elements differ from the background (non-regulatory) DNA. There seem to be three key points defining the regulatory elements:

1. The regulatory elements are conserved in evolution.
2. The regulatory elements occur in more than one location in the genome.
3. The regulatory elements near a gene affect the expression of the gene.

All of these should be taken into account when discovering regulatory elements in the genome. However, one can start from any of these key points by generating the candidate list for regulatory elements and then filter out those candidates that do not satisfy the other two conditions. Let us describe one possible method starting from each of these key points.

1. One possibility of detecting evolutionally conserved regions is using alignment. Once found, it is possible to check if these occur in more than one location in the genome (with approximate matching) and if they affect the expression of the gene — the latter can be done by comparing the expression profile of all of the genes this regulatory element is close to. The weakness of this approach is that it requires alignment, which is difficult to obtain in the case of very divergent species.
2. The regions which are repeatedly occurring in the genome can for example be found by counting the k -mers (substrings of length k). Those that occur

significantly more often than expected, are probably parts of regulatory elements. The resulting candidate elements can be checked for the other two key points. The weakness here is that the regulatory elements have to be occurring very often to be discovered.

3. The genes with similar expression profile are supposed to have at least partly the same regulatory elements in their promoters. Although we did not get an explicit candidate element set yet, we can now try to find the substrings that occur in the promoters of similarly expressed genes more often than in the promoters of other genes. These substrings are the candidate elements and can be checked for evolutionary conservation. The weakness of this method is that the expression levels have complicated patterns and therefore, genes with mostly similar but still slightly different promoter regions can have totally different expression profiles.

This concludes our overview of possible approaches to the problem and now we concentrate on one specific method.

3 Method by Elemento and Tavazoie

Elemento and Tavazoie (Elemento & Tavazoie 2005) go in the direction of the first approach (see above), but they try to avoid the weaknesses of alignment-based approaches. Instead of producing a multiple alignment for several organisms, they just take two organisms (which are allowed to be very divergent) and find k -mers occurring in both organisms. This produces a huge list of candidate k -mers. Next they make use of the mixture of first and second key points, and rank these candidates by calculating the number of orthologous genes in the two organisms that have the k -mer evolutionally conserved. Here the actual ranking is not produced based on the absolute number of orthologous genes but based on the probability that this number would be so large compared to the number of total occurrences of the k -mer in the random setting with hypergeometric distribution. Putting all this into formulae:

- Take a list of orthologous gene pairs $(G_1, G'_1), \dots, (G_N, G'_N)$ from two organisms O and O' as input to the method (usually obtained from some public database).

- For each k -mer ($k = 7, 8, 9$) and for each organism (O and O') compile a set of genes which has this k -mer in the promoter, let these sets be S and S' for the two organisms, respectively.
- Let the sizes of sets be $|S| = m$ and $|S'| = m'$ and their overlap $|S \cap S'| = r$. The probability of two sets of size m and m' , drawn from a set of N elements, to have r or more elements in common is given by:

$$P(X \geq r) = \sum_{i=r}^{\min(m,m')} \frac{\binom{m}{i} \cdot \binom{N-m}{m'-i}}{\binom{N}{m'}}$$

Smaller probability corresponds to higher ranking of the k -mer.

Now they have got a ranked list of k -mers and as a last step they throw out the k -mers that have the probability value higher (and ranking therefore lower) than some threshold. This threshold is the only parameter to the method.

The strength of this method is that it is easy to incorporate any kind of biological data which is representable as a gene set. This includes the gene ontology (gene categories based on their molecular function, cellular component, or process they are involved in), but also the gene expression data (e.g. set of genes with similar expression profile). The gene set X can be used by checking the significance of the overlap of sets X and $S \cap S'$. If the overlap is significant, then the k -mer is relevant with respect to the gene set X .

The weakness of this method is that it depends on k -mers, which are exact matches of the substring. However, in reality the approximate matches are also good enough for binding.

4 Summary

In this short paper we have first given three key points to in silico discovery of regulatory elements in genomes. These lead to different methods, a couple of which we have shortly described. We have gone into more detail describing the method by Elemento and Tavazoie (Elemento & Tavazoie 2005), which is one of the most recent ones in this field. The comparison of the results obtained from different methods is beyond the scope of this paper.

References

Elemento, O., and Tavazoie, S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology* 6(2):R18.