Using sets of orthologous genes to discover regulatory sites in DNA sequences.

Dieter De Witte

Department of Information Technology, Ghent University, VIB, N2N

Supervisor(s): Jan Fostier, Bart Dhoedt

Proteins take the lead part in biochemical processes in all living beings. The information to build proteins is stored in the genes. A gene is some sort of a blueprint to build a protein. The process to generate the proteins from the DNA is shared between all organisms: A gene is read from the genome and copied into a mobile RNA molecule, which can leave the cell's nucleus. This is called transcription. In a next step this RNA molecule is read by a ribosome and translated into a protein.

In human most genes have been identified. This doesn't mean the job is done. Genes only make up for about 1 percent of the total genome. The remaining part is called junk DNA. Originally scientists thought this junk DNA didn't have any biological function. Later on it was discovered that gene expression - the rate at which a gene is translated into a protein - is influenced by proteins binding in the neighbourhood of the gene, in the junk DNA. The proteins connect to the transcription machinery and can increase or decrease the transcription rate. The binding sites can be interpreted as switches to control our transcription.

The identification of these binding sites could have many applications. It might give us the knowledge to devise pharmaceutical drugs for diseases which have a genetic origin.

Discovering the actual binding sites for the regulatory proteins is a very challenging job. First of all it has been shown that the binding site does not have to be an exact DNA word in order to have a succesful binding. The protein may also bind if a letter is missing or it might allow two different letters at a given position. This implies that if we look for a consensus representation for those binding sites, we have to look for words that match approximately.

In this research IUPAC degenerate strings are used. This are words over an extended alphabet containing degenerate symbols matching with two or more characters.

Since the DNA alphabet contains only four characters it is very difficult to judge whether a certain word in a DNA string is simply a lucky shot (= false positive) or it is a good candidate binding site. One idea is to look for words which occur alot of times. Unfortunately this criterium has proven to be unreliable. In this research we study groups of related genes in different organisms, called orthologs. Since the gene came into existance in an ancestor of the organisms we expect the binding sites regulating it to be conserved as well. We are therefore looking for words which are surprisingly conserved within a family of organisms. This approach allows us to come up with a statistical procedure that is able to make more reliable predictions.

We have developed a fast algorithm to find words in a DNA text using an index structure called a generalized suffix tree. We have also distributed our software in order to make it work on the university's supercomputer. Currently data from a number of plant datasets (monocot and dicot families) is being prepared at the Flanders Institute of Biotechnology (VIB) to be investigated with our software.