7. Introduction

In eukaryotes the promoter region of genes can contain different regulatory elements. These elements are capable of binding transcription factors. Some transcription factors could be conserved among orthologous species. By use of multiple sequence alignment these conserved motifs can be found with bioinformatic methods.

Previously our group had created a database called DoOP to collect the promoter sequences from plants and chordates. This database contains thousands of conserved motifs, but the connection between the motifs and the biological function of the regulated genes was unknown.

The degree of conservation can be determined at different levels. In the case of chordata, conservation in *Primates* is not always found in the *Mammalian* group. For this reason DoOP contains different conservation levels called subsets.

Motifs usually do not possess an exactly identical nucleotide sequences. A number of methods can be used to describe nucleotide variations in a given position of the motifs. The DoOP database used consensus sequences in which the variations were marked as IUPAC codes. Lower- and uppercases also have different meanings. Lowercase denotes the less feasible nucleotides. Most of the programs use a matrix representation for the comparative analysis of motifs. It was necessary to develop a new program for comparing of consensus sequences.

Expression arrays can identify large number of genes for which transcription was induced by the same biological processes.

ChIP (chromatin immunoprecipitation on chip) can determine the location of DNA binding sites on the genome. This work based on chip experiments. Expression array was used to find genes of a given biological effect and ChIP was used to find histone modification positions in promoter regions.

Gene ontology (GO) is the bridge between the gene and its function. It is a hierarchical database where the parent nodes always contain general functionalities and the child nodes are more specific. Searching overrepresented gene ontology terms in large sets of genes, like array experiments, is a common method.

8. Aims

The main aim of this study was to find connection between the nucleotide sequence of the motifs and the biological role of the regulated genes. In order to reach this goal we needed to develop a new algorithm which compared motifs with each other, and to add new features to the DoOP database. Using this algorithm similar motifs were collected and we tried to find identical functionalities of the genes belonging to the given motifs by gene ontology. Identical motifs were searched for in promoters of the coexpressed genes derived from microarray experiments. Finally histone modification regions were analysed to find motifs which were specific for a biological pathway.

9. Materials and methods

9.1. Bioinformatic databases

The primary source of DNA sequences was the EnsEMBL genome database. To fetch the sequences the EnsEMBL Perl Application Programming Interface was used. The source of promoter homology experiments was the EnsEMBL Compara database or the UCSC database. The Gene Ontology database was used for determining the gene functions (released in february, 2007).

9.2. Sequence processing programs

The search of simple consensus sequences in large number of sequence files was performed by the *fuzznuc* program from Emboss 3.1. Two missmatches were enabled in the course of these runs. Generating the control dataset's original sequences was mixed with *shuffleseq* program, so nucleotide ratios of the control datasets ranged with the original ones.

Repetitive sequences were removed by censor 4.1. In the case of human datasets the *-hum* command line parameter was applied. This setting removed not only the human specific repeats but also the ALU sequences.

9.3. De novo motif discovery

New motifs were searched for by NestedMica 0.7.3. This program was very sensitive to the *-ensembleSize* parameter which was set to a value of 400 based on test runnings. The length of the target sequence was 13 basepairs. The program can find a given number of motifs which can be set by *-numMotifs* argument. In this experiment the adjusted number of motifs was 4.

NestedMica is capable of finding conserved motifs in the orthologous sequences. In this experiment *Canis familiaris*, *Bos taurus*, *Mus musculus* and *Rattun norvegicus* were applied.

9.4. Processing chip and chip-on-chip results

Expression array experiments with Affimetrix HG-U133A plate were made by István Szatmári. Analysis was carried out by means of GeneSpring 7.3.1 software and Bioconductor 2.0. The raw data were analyzed by the GC-RMA algorithm then results were normalized using per-chip normalization. Probe sets, which had lower expression levels than 20 in 90% of the experiments, were filtered. Two-fold up- or down-regulated probe sets (p ¡ 0.01) were selected.

Chromatin immunoprecipitation was carried out by László B. Bálint. The HL60 cell line was treated with retinoid acid to start differentiation. Control cells without treatment were marked as naive. In the immunoprecipitation step two types of antibodies were used. The first one was specific for the histone H3K4 chain and the second one for the histone H4R3 chain.

The raw data was processed by Affymetrix Tiling Analysis SDK 2 program with the following settings: type: 0, band: 25, pval_scale 0, sig_scale 2. The sequences attached these positions were extracted from NCBI database version 35.

9.5. Statistical analysis

Hypotheses were checked with the R package (version number: 2.4). All the tests were estimated to have a probability of error less than 0.05. Averages were compared with *t.test* program and correlations were examinated by the program *cor*.

To find overrepresented gene ontology terms a modified GeneMerge program was applied. This program uses a hypergeometric distribution based model to find significant GO terms. GeneMerge was modified because the original program had stored intermediate data in files instead of memory.

10. Results and discussion

10.1. New DoOP database features

The main new feature of the DoOP database is the programming interface (API). Using this API adding new features to the website is much easier. Its advantage is that data can be fetched from command line scripts. Most of the bioinformaticians use Perl for daily programming tasks, for this reason the DoOP API was implemented in this language.

This programming interface was published in CPAN, which is a large collection of Perl software. It is completely object oriented. To avoid module name collisions it has its own namespace called Bio::DOOP:: DOOP. There are no Bioperl dependecies in this module.

The API contains helper classes to run motif comparisons and produce PNG output from the positions of the motifs.

10.2. The motif comparison algorithm

Mofext was developed to determine motif similarity. This program cut the query and target sequences into smaller overlapped words. The words were compared by a comparison matrix. This matrix contained the similarity scores of the bases. If similarity of words was greater than a cut-off value then the program tried to extend the words. Finally, we get the extended subsequences and the similarity score. *Mofext* can handle only gapless motifs.

A DoOP module can call this program. This feature helps users to fetch motifs from the database and run similarity searches.

To test the abilities of the program four query motifs were selected from Transfac database. These motifs regulated genes which took part in cell cycle regulation, homeostasis maintain, and neuron- and muscular development. The targets were the motifs in the the DoOP chordata.

The gene ontology of the similar target motifs was analysed with a text based process. To match gene ontology to a motif the gene ontology of the gene was used which can be found towards the 3'-end of the motif. The gene ontology terms were searched for by keywords, because the Transfac database at that time did not contain gene ontology terms.

The results showed that increasing the wordsize parameter for *mofext* found more specific motifs. However, gene ontology has some imperfections. For example not all investigated motifs could be matched to

ontology, because not all genes had this type of features. Another problem was that the structure of the gene ontology was a directed acyclic graph. To find all GO terms that fit to a gene, it need to check all the nodes which requires intensive computer work. In this investigation only the last GO nodes were searched for the term.

10.3. Motif clustering

Mofext is able to compare motifs, so it can be use to cluster motifs. A new program was developed to compare motif groups called *klaszterezo.pl*. This program can run in parallel mode.

Some of the investigated motifs came from chromatin immunoprecipitation experiments. Histone acetylation and metylation associated positions from the ENCODE region were extracted and homologous sequences were found with the help of UCSC Genome Browser.

Homologous sequences were aligned with the program Dialign2. Conserved segments were picked from these multiple sequence alignments and 4708 motifs were found.

After clustering these motifs were ranged into 21 classes. To find out what is the identical biological function of one cluster GO was used. Overrepresented gene ontology terms were searched for using GeneMerge. None of GO categories were characteristic of only one cluster. It may be that too many false positive motifs were in the clusters.

Following this we had tried to cluster all the motifs from DoOP database, which would take 113 years on the basis of preliminary calculations. To decrease the running time indexes were used. However, indexing of 4 million motifs would take up too much space, therefore only chordata subsets were indexed and clustered. Results showed that closing to the *Primates* subset the average number and length of motifs in a sequence alignment increased.

Mofext splits the motifs into smaller chunks therefore the triangle inequality was sometimes broken. In this case the clustering process did not produce discrete clusters but motif gradients. This behaviour depends on the word size of *mofext*.

All these results suggest that Mofext is good for finding similar motifs but it makes an error in case of clustering. Further development needed to refine the weaknesses of this program.

Motif clustering is not a solved task. The main problem is to compare motifs with a different length. Developers of the cisRed database also tried clustering motifs. They had very large computer capacity, but they also failed when they tried to compare motifs with different length.

10.4. De novo motif discovery

To find new motifs a special biological process was chosen. This process was the monocyte-dentritic cell differentiation. The differentiation was induced by rosiglitazon. Samples were taken 6 hours, 24 hours and 5 days after the beginning of the differentiation. Upregulated and downregulated genes were chosen from every sample.

Promoter regions of the genes were extracted from the EnsEMBL database. Previous articles suggested that the lipid-regulated nuclear receptor peroxisome proliferator-activated receptor gamma (PPAR*gamma*) is involved in the dendritic cell development.

Canonical PPAR*gamma* binding sites (AGSTCMN(1,7)AGSTCM) were searched in the investigated and the control gene lists. Two types of control gene lists were used. The first one contained randomly chosen genes matching the number of the original dataset. The second one contained the original genes of which sequences were shuffled.

Comparison of the different sets showed that the shuffled dataset always differed from the other two datasets, but there were no differences between the original and randomly chosen datasets. These results suggested that the PPAR*gamma* is a more general transcription regulator than it was presumed beforehand.

Specific transcription factor binding sites were searched for by NestedMica. First executions did not produce any notable results. NestedMica can use orthologous sequences to find more specific motifs. Orthologous sequences were retrieved from EnsEMBL database. Using these sets NestedMica found motifs in promoters of upregulated genes in samples that were taken 6 hours after the start of the differentiation. The consensus sequence of the most prominent motif was RCCTCNRCCTC. It had a heterodimer form the same as PPAR*gamma*, so it was called DRA in this phase of the work.

The presence of DRA was counted in the original dataset, in the randomly chosen promoters and in the shuffled sequences. Comparison of the sets showed the same results as in the case of PPAR*gamma*. The largest amount of DRA was found near to the TSS. Unfortunately if the repeats were removed with RepeatMasker software instead of Censor, the DRA disappeared around the TSS. Further investigations are needed to decide what its origin is.

10.5. Chromatin immunoprecipitation

The HL60 myeloid cell line differentiation was started by DMSO, but it was induced by 9-cis retinoid acid. During the differentiation using chromatin immunoprecipitation against histone modification factors the positions of the acetylation and methylation points could be identified.

The distance of the acetylation and methylation from the nearest TSS was measured and counted in every 200 base pair segments. These segments overlapped by 50 base pairs. The number of the presence of these histone modification points was drawn as a graph and it was established that the distribution from the TSS is systematic.

At this point a novel method was developed that be able to filter ChIP results with expression array outcomes. We had chip results by experiments on HL60 cell treated with 9-cis-retinoid acid. Gene lists were created from expression array and ChIP experiments as well. The gene with correct orientation which was found nearest to the TSS was matched to the ChIP positions. The ChIP results could be hereby filtered with the up-, down- and unregulated genes. The TSS dependency effect appeared in all cases.

Some scientific publications have suggested that the exon-intron junctions have special role in gene regulation therefore the distance of the acetylation and methylation positions to the exon-intron junctions was also measured.

The results showed that the number of the histone modification positions was the largest in junctions and they occured in greater number in the introns than in the exons. The previously mentioned filtering did not indicate any differences. On the basis of our results we can state that the exon-intron junctions have an important role in gene regulation but they did not show any pathway-specific differences.

11. Summary

Finding connections between the gene function and consensus sequence of TFBS and its promoter region is not an easy task. First of all, a new bioinformatic background was needed. It was achived by the creation of the DoOP database, the `mofext` program and by the combination of these two in the DoOPSearch webpage. These tools are not specific for a certain investigation, they can be used for any other promoter and conserved motif research.

As a second step, we collected motifs that were similar to each other and tried to search for common functions in their genes. We have used a gene ontology database to identify the common biological functions of the genes and a method based on hypergeometric distribution was used to determine the percentage of similarity. Clustering of these groups was beyond our resources, so we have used it only in groups of animals that are a small evolutionary distance from each other.

For the sake of the cause we reversed the methods. Instead of collecting genes with similar motifs, we tried to search common control elements in the genes with similar biological function. Gene collections were obtained from chip experiments. All the collected genes contained PPARγ, a transcription factor binding site that proved to participate in lipid metabolism. A new method was developed to analyze the genes with positive responses in chip experiments. We demonstrated that the promoters of overrepresented genes did not contain statistically more DR1 than the randomly selected promoters. With the help of *de novo* motif searching, new elements were identified from the gene list. However, their biological role is still not clear.

Analysis of the distance between the transcription start site and the motifs was more succesful. It was shown that the probability of occurence of the motifs is increased nearby the TS-Sites. These results suggests that the regulation depends not only on the presence or absence of the motif, but also on its position in the promoter. Other publications also support this suggestion.

Another indirect indication of the position specific presence can be the relationship between the positions of the exon-intron junctions and the acetilation or methilation points. This distance between the maximal number of junctions and the acetilation and methilation points proved to be 154 basepairs. Without yet understanding the nature of the relation between the first base of the intron and the acetilation and metilation points, it can still be concluded that the linkage depends on their distance from each other.

The complex picture of transcription regulation is far from completely understood. For example it is still unclear how far the boundaries of the promoters extend or whether there is any genomic structure that can also influence the gene regulation.

It is shown that the topology of DNA can also be conserved. Different sequences can produce similar three-dimension topologies that can be recognized by elements of the regulation. These pieces of information should also be integrated into the bioinformatic analysis to get more accurate results.