

# Fine scale regulatory annotation of a gene

Katalin Orosz  
Eötvös University of Sciences

August 14, 2008

## Contents

1	Introduction	1
1.1	ORMDL genes	1
1.2	Eukaryotic gene structure	2
2	Methods	3
2.1	Comparative genomics	3
2.1.1	Comparative methods	3
2.2	Methods based on a priori knowledge	5
2.3	Collecting sequences from the databases	5
3	Results	6
3.1	Initial analysis of the sequences	6
3.2	Comparative methods	9
3.3	Known motifs found	10
4	Conclusion and future directions	14
5	Acknowledgement	17

## 1 Introduction

### 1.1 ORMDL genes

ORMDL1 gene was characterized from retinitis pigmentosa. Initially, Human ORMDL1 cDNA, then ORMDL2 and ORMDL3 were characterized and it was revealed that ORMDL1 belongs to an evolutionary conserved gene family. In vertebrates, three different genes were distinguished, corresponding to

ORMDL1, ORMDL2 and ORMDL3. Homologs for the three ORMDLs were found in yeast, microsporidia, plants, Drosophila, urochordates and vertebrates. (Hjelmqvist et al., 2002)

As earlier studies discovered there are some SNPs (Single Nucleotide Polymorphism) compared to healthy people in ORMDL genes that can be associated with asthma. The genome-wide association of SNPs with childhood asthma showed that there are multiple markers on chromosome 17q21 associated with this disease. The SNPs were also strongly associated with the transcription levels of ORMDL3 (Moffatt et al., 2007). Therefore, it is clear that understanding the regulatory mechanism of ORMDL genes can provide us some really valuable information about the appearance of asthma and the operation of genes responsible for it.

The main goal of the project is to annotate the ORMDL genes in Human. We tried to use some techniques of comparative genomics and regulatory motif identification in order to find TFBSs (Transcription Factor Binding Sites) in the ORMDL genes. The three ORMDL paralogs in Human are ORMDL1, ORMDL2 and ORMDL3, which are located on different chromosomes. The genes are highly similar but their regulatory region may vary and their function is still unknown.

After collecting genomic and proteomic data from some different species we tried to use sequence data for getting some novel information about the upstream regions, in which we are expecting to find TFBSs.

## 1.2 Eukaryotic gene structure

Genes usually have a transcriptional and a regulatory region. Transcriptional region is transcribed into primary transcript and after further processing it is translated into a protein. Regulatory regions usually consist of cis-regulatory elements and trans-regulatory elements. Trans-regulatory elements are DNA sequences which encode transcription factors. Cis-regulatory elements are binding sites of the different kinds of transcription factors. Regulation is done by proteins binding to these elements and affecting the transcription of the gene. The promoter region is responsible for transcription initiation. When a transcription factor binds to an enhancer region it enhances transcription. On the other hand, binding of transcription factors to silencers tend to repress transcription. At last but not least there are response elements, which are recognition sites of transcription factors. Many eukaryotic genes have a so-called TATA box. This lies usually very close to the transcriptional start site, often within 50 bases. A TATA box can bind a TATA binding protein, which assists in the formation of the RNA polymerase transcriptional complex.

## 2 Methods

### 2.1 Comparative genomics

One of the approaches for finding regulatory regions is to use comparative methods, where they are searching for highly conserved elements in orthologous sequences. Another way to predict TFBSs is to search for already known motifs in each specie. The still exponentially growing genome sequencing activity has opened the door for researches that would use the genome or protein sequence information to perform further analysis on them. As there are more and more species with genome sequence revealed we can use the sequence information of these species for comparative approaches. Comparative genomics can be a very useful approach which helps us understand the structure and function of genes. It can provide some results that can either give idea for new strategies against diseases or give us some novel information on evolutionary changes. Comparative genomics studies focus on the relationship between the genomes of different species. It tries to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that can act on genomes. Recently it focuses on finding genes and non-coding functional elements of the genome. The approach is based on looking for similarities in proteins, RNA and regulatory regions among different species. Elements that are responsible for functionality should be conserved throughout the time, however elements without any functional role tend to be divergent. Therefore, looking for conserved or slowly evolving elements is a common way of revealing regulatory elements.

#### 2.1.1 Comparative methods

We applied FootPrinter, DiAlign and SAPF for the collected sequence data, see subsection 2.3.

FootPrinter ([Blanchette and Tompa, 2003](#)) finds highly conserved regions of sequences which are hence predicted as regulatory elements. The algorithm solves the Substring Parsimony Problem which, on one hand, is NP-hard, however, on the other hand, an exact solution using dynamic programming exists that is sufficiently fast in practice.

As input data FootPrinter needs orthologous sequences  $(S_1, S_2, \dots, S_n)$  and the phylogenetic tree joining them ( $T$ ). The problem is to find all sets of substrings  $s_1, s_2, \dots, s_n$  of  $S_1, S_2, \dots, S_n$ , each of length  $k$  while the parsimony score of  $s_1, s_2, \dots, s_n$  on  $T$  is at most  $d$ . (The parsimony score of a set of sequences is the total number of substitutions over the tree  $T$ .)

The dynamic programming algorithm assumes a rooted tree (the tree can be rooted in an arbitrary internal node,  $r$ ). It computes  $W_u[s]$ , which means for a string  $s$  of length  $k$  the best parsimony score that can be achieved for the subtree rooted at  $u$  if  $u$  was to be labeled with  $s$ . ( $u$  is a node of the tree.)  $W[s]$  consists of  $4^k$  entries as there are  $4^k$  different nucleotide sequences of the length  $k$ .  $W$  can be computed by a dynamic programming algorithm, which proceeds from the leaves up to the root. The solution for the Substring Parsimony Problem is a minimization:

$$\min_{s \in \Sigma^k} (W_r[s])$$

which means the minimization of  $W[r]$  for all possible values of  $s$ .  $\Sigma = \{A, C, G, T\}$ . Then substring  $s_1, s_2, \dots, s_n$  can be recovered by tracing back the recursion, from the root to the leaves, for each entry of  $W_r$  with score at most  $d$ .

DiAlign (Morgenstern, 1999) performs a local alignment on sequences from multiple species therefore, is able to find isolated regions of local similarity. First of all, it finds gap-free pairs of sequence segments of equal length from pairwise alignments, so-called diagonals. The next step is to choose a consistent set of diagonals with a maximum sum of weight. (Consistent means that all segment pairs are matched in an alignment.) To get the weight of a diagonal a similarity value is defined,  $w$ . In case of each diagonal for each nucleotide residue pair  $w = 1$  if they are a match and  $w = 0$  otherwise.  $s_D$  is the sum of the individual similarity values of residue pairs within the given diagonal. The weight of a diagonal in DiAlign 2 is computed according to the following expression:

$$w_2(D) = -\log P_2(l_D, s_D)$$

where  $P_2(l_D, s_D)$  is the probability of finding any diagonal of length  $l_D$  for which the sum of the individual similarity values is at least as large as  $s_D$  somewhere within the comparison matrix of two random sequences with the same length as the original sequences.

The consistent set of diagonals are collected into  $M_1$ . After the calculation of the overlap scores of the diagonals in  $M_1$  the scores of the diagonals are updated. The sorting of the items in  $M_1$  according to the updated scores starting from the largest score is followed by putting them into  $M_2$  provided they are consistent with the previous ones. In the next iteration step the sequences that were not directly aligned are aligned and the same procedure is repeated for them.

The input file should contain the nucleotide sequences in fasta format. The output file contains the aligned sequences. In each column of the align-

ment a number from 0 to 9 indicates the similarity of the sequences, where 9 is used for the most similar regions.

SAPF (Satija et al., 2008) combines the features of a statistical aligner and a phylogenetic footprinter. The algorithm does the multiple alignment and the annotation of the input sequences based on a hidden markov model. Each column of the aligned sequences is annotated as functional or neutral. Phylogenetic footprinting is carried out by defining slow states and fast states which means that the number of the states used in the HMM must be doubled. A set of states are fast states and the other set of states are slow states. For each alignment column, the computed posterior probability shows the probability that the nucleotides in the particular position are homologous and should be aligned in a single column. The probability of being in a fast or a slow state is also computed. Slow states correspond to slower divergence, which is the result of a purifying selection and characteristic for the regulatory regions. Fast states correspond to a higher level of divergence being characteristic for neutral sequences.

## 2.2 Methods based on a priori knowledge

There is another way of finding regulatory elements besides the previous ones based on highly conserved regions. There are some databases containing information about the TFBSs, which can be used for searching for known motifs in our sequences. One of these databases is TRANSFAC. (Matys et al., 2006) TRANSFAC is the database that collects data which are relevant for gene expression at the transcriptional level. Transcription factors interact with short, 5-25 nucleotide long DNA elements the TFBSs. Therefore, the basic information that can be found in the tables are collected in SITES and FACTORS. (Wingender et al., 1996)

As a motif finding tool we used MotifScanner for the identification of possible TFBSs in our nucleotide sequences. In the given sequences MotifScanner (Aerts et al., 2003) finds the possible transcription factor binding sites based on a given motif model and background model.

## 2.3 Collecting sequences from the databases

We collected nucleotide and protein sequences from different species to be able to analyse ORMDL genes from different aspects and get as many information about the features of them as possible.

Initially Ensemble seemed to be a practical choice for deriving sequence information, however at the end the most complete sets of sequences were

available from the NCBI database. The contig information contained relatively complete sequences. (Incomplete means that in some regions of the sequences there were unknown nucleotides.) We were also interested in the additional information on the mRNA, CDS regions and splicing variants. In case of the protein sequences, there were no gaps, it was possible to get complete sequence data.

For ORMDL1, it was possible to find the homologous sequences only in Human, Rhesus, Mouse. For ORMDL2 and ORMDL3 five homologous sequences were collected: Human, Chimp, Rhesus, Mouse, Rat. It would be a reasonable question why these sequences are in the focus of our investigation. It was obvious from previous analysis of the data and using the UCSC genome browser to look at homologous sequence diagrams, that choosing two more distant species, for example Human and a fish some longer regions totally disappeared from the fish. Therefore, we decided to download the sequences only for some less distant species.

TFBSs detection requires 5' upstream regions for detailed analysis. In order to get the proper regions firstly all nucleotides from the 5' end to the 3' end of the mRNA and an additional 5000 nucleotides on both ends were manually derived.

As a next step, we determined the range that was to get from the sequences and cut 1000 nucleotides from the upstream region in front of the 5' end of the mRNA. After cutting the proper sequence parts in many cases the regions had to be complemented. Protein sequences were available for each mRNA from NCBI. They usually consist of 153 amino-acids but in two cases, namely at Rhesus and Rat in ORMDL3. When performing initial analysis of the sequences it was usually practical to use such sequences that are of equal length, therefore the first couple of nucleotides in case of which there were no match with the other ORMDL sequences were ignored.

## 3 Results

### 3.1 Initial analysis of the sequences

The downloaded and processed sequences had to undergo some kind of rough checking procedure which proves that they are homologous and are roughly related in a phylogenetic tree according to our first expectation. (It means that for example a nucleotide sequence from Human is supposed to be closer to the Chimp than to the Mouse.) As a rough estimation of the sequence quality, ClustalX ([Chenna et al., 2003](#)) was used to perform an alignment for the entire ORMDL3 sequences (the mRNA region with the 5000 nucleotide

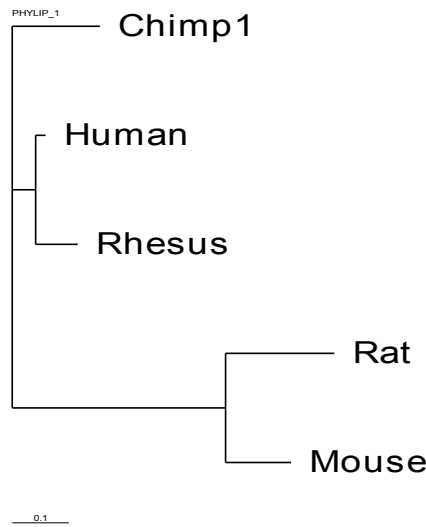


Figure 1: The neighbour joining tree of the complete ORMDL3 sequences created by ClustalX. The relationship between species roughly corresponds to the expectations.

long upstream region and the downstream region of the same length). For these we got positive results. (Figure 1.) The neighbour joining tree roughly corresponds to our expectation.

For TFBS identification the 1000 nucleotide long upstream regions are required. For each ORMDL nucleotide dataset, with 1000 5' end nucleotides, a ClustalX alignment was done, which did not provide so convincing result as regards the relationship of the species and the ratio of matched sequence parts in the alignment. (Figure 2) Some more unexpected relationships among species also appeared.

However, the neighbour joining tree of ORMDL1 is obviously less infor-

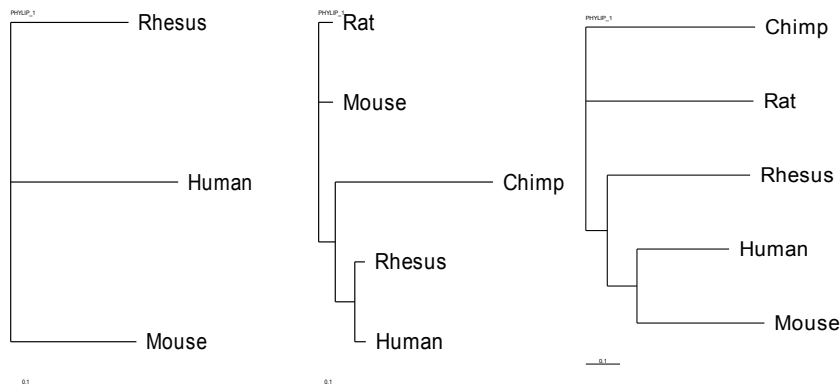


Figure 2: The neighbour joining trees for the ORMDL1, ORMDL2 and ORMDL3 upstream sequences created by ClustalX.

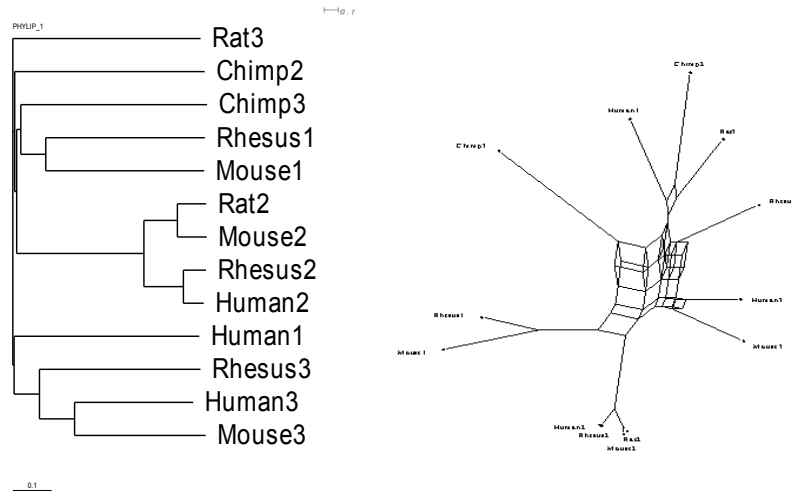


Figure 3: The neighbour joining trees for the merged dataset, which consists of the ORMDL1, ORMDL2 and ORMDL3 upstream nucleotide sequences, created by ClustalX (left). The consensus network of the upstream nucleotide sequences in the merged dataset based on the calculation of MrBayes (right).

matic because it contains only three species. The neighbour joining tree and the matches for ORMDL3 were really bad. The sequence analysis with FootPrinter and DiAlign, which is described more detailed later, referred to the different structure of ORMDL2 Chimp nucleotide sequence. These points made us performing a short analysis of the nucleotide sequences for all three ORMDL genes together. All these sequences were put into a common file and a ClustalX alignment was performed to this merged dataset. The results were given as an input to MrBayes, which approximates posterior probability distribution of trees with Markov Chain Monte Carlo. We created a consensus network using SplitsTree4 (Huson and Bryant, 2006). The consensus network shows all possible trees which appeared at least in 10% in the sampling of the Markov Chain. The only valuable information that was obvious from the network is that ORMDL2 Chimp sequence is separated from the other ORMDL2 sequences. It suggested that ORMDL2 Chimp sequence might have a wrong annotation. (Figure 3)

After merging the protein sequences of equal length another ClustalX alignment and a MrBayes analysis revealed that the different ORMDL genes are separated on the plot, which means that they form paralogous groups with orthologs in each group. The evolution of ORMDL1 was the slowest one among the three paralogs and the division of the Human, Rhesus and Mouse shows uncertainty. However, there is a well-defined division of the Primates and the Rodents in case of the ORMDL2 and ORMDL3. (Figure 4)

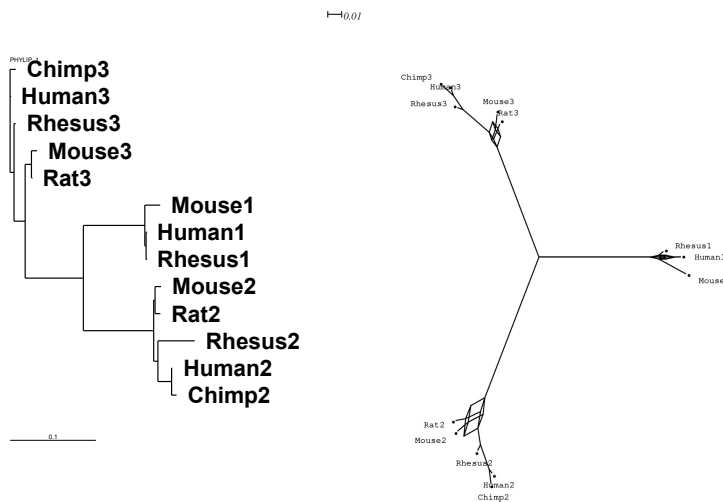


Figure 4: The neighbour joining trees for the merged protein dataset, which consists of the ORMDL1, ORMDL2 and ORMDL3 protein sequences, created by ClustalX (left). Since the tree is not a rooted tree, it is possible that the ORMDL3 sequences are not the most ancient ones, contrary to the suggestion of the picture. The consensus network of the protein sequences in the merged dataset based on the calculation of MrBayes (right).

### 3.2 Comparative methods

FootPrinter generally found no motifs or too many motifs for the ORMDL nucleotide homologs. The best results were obtained for ORMDL2, which contained less repetitive sequence parts than the other ORMDL sequences. That is why FootPrinter was able to identify some consistent series of conserved regions in ORMDL2 but did not perform any meaningful result for the other ORMDLs. Moreover, FootPrinter found many inconsistent motifs besides the consistent ones among the found conserved regions as well. A possible explanation of the results can be that FootPrinter looks for relatively short regions of conservations with too few number of mutations. That is why FootPrinter may be unable to find the series of conserved regions of different size separated by short nonconserved regions. (Figure 5)

The alignment method of DiAlign is based on joining similar regions together in a consistent way. This method is able to cope with local similarities and even find blocks of highly similar regions. The DiAlign alignment was done for all the sets of ORMDL sequences. DiAlign similarity regions with similarity score from 6 to 9 were collected from the alignment data and each similarity region was drawn onto a plot with a Perl script. On each plot the similarities found by DiAlign are illustrated with boxes of different colours.

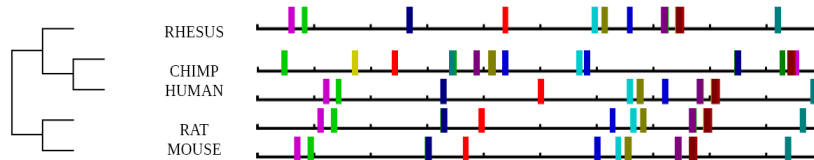


Figure 5: Found conserved regions (motifs) with FootPrinter for ORMDL2 upstream nucleotide sequences. Different colours indicate different motifs. The phylogenetic tree next to the motifs was provided by FootPrinter. There are consistent motifs, which means that the same series of these motifs can be found in all five ORMDL2 sequences. However, there are many artifacts as well, that cannot be highly conserved sequence part. (FootPrinter 3.0, motif size: 11, maximum number of mutations: 3, maximum number of mutations per branch: 2)

Each colour corresponds to a level of similarity. The pictures illustrate how these boxes are arranged along each 1000 nucleotide long upstream sequence, probable insertions and deletions can be easily detected. In case of ORMDL2 there are many blocks of simialrity regions wich are separated with shorter less similar regions. The sequence of Chimp shows different pattern compared to the other species. This refers to possible intensive indel events in the regulatory region in this case. In ORMDL1 the corresponding similarity regions can be found in all the three genes but some insertions also can be supposed. (Figure 6) The ORMDL3 Rat and Mouse sequences show few or no similarity regions. That can happen if the upstream regions does not contain the same regulatory regions as the Human, Chimp and Rhesus, so we maybe failed to get the proper upstream nucleotides for Mouse and Rat. (Figure 7) The Human sequence seems to be somehow 'shifted' compared to the others, i.e. if we leave the last 500 nucleotides from the end of the 1000 nucleotide long upstream region (3' end) and get 500 nucleotides towards the beginning we may get those parts that were found in the Chimp and the Rhesus. Therefore, we created a dataset which contains the Human sequence from the 501st nucleotide in front of the mRNA to the nucleotide position 1500, towards the 5' end. After the DiAlign alignment the similarity regions inferred the hypothesis, that probably some insertion occured in the Human sequence.(Figure 8)

A fast run of SAPF found three regions that can be transcription factor binding sites. (Table 1)

### 3.3 Known motifs found

The sets of ORMDL sequences (ORMDL1, ORMDL2, ORMDL3 and the ORMDL3 dataset with the reprocessed Human sequence and the original

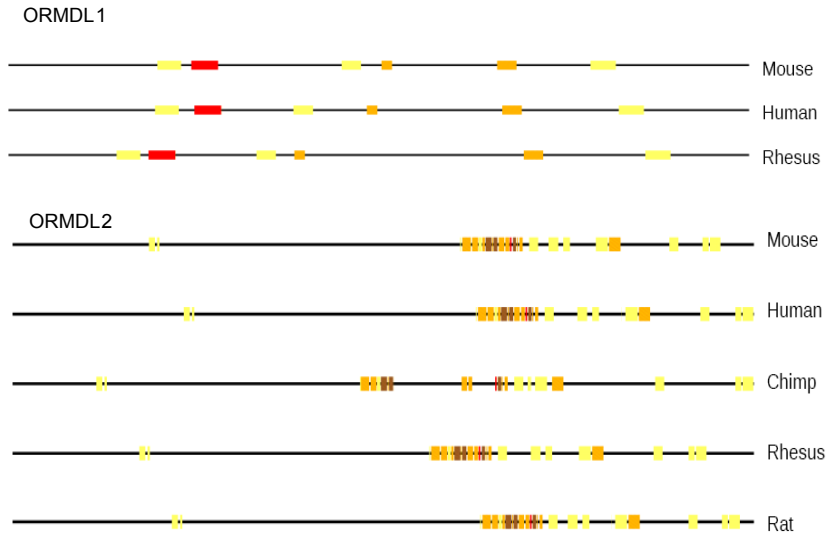


Figure 6: The highly similar regions according to the DiAlign alignment for ORM DL1 and ORM DL2. Regions with the highest similarity are indicated by red (similarity score 9), then 8 is brown, 7 is orange and 6 is yellow.



Figure 7: The highly similar regions according to the DiAlign alignment for ORM DL3. Regions with the highest similarity are indicated by red (similarity score 9), then 8 is brown, 7 is orange and 6 is yellow.



Figure 8: The highly similar regions according to the DiAlign alignment for reprocessed Human, original Chimp and Rhesus ORM DL3. Regions with the highest similarity are indicated by red (similarity score 9), then 8 is brown, 7 is orange and 6 is yellow.

Human	ctctgggggtcccactccagc	674..694
Rhesus	ctctgggggtcccactccagc	853..873
Chimp	ctctgggggtcttctctcaagc	848..868
Human	gaggatg	339..345
Rhesus	gaggatg	520..526
Chimp	ggggatg	526..532
Human	ggggagg	425..431
Rhesus	ggggagg	604..610
Chimp	ggggagg	612..618

Table 1: Found motifs for ORMDL3 reprocessed Human, original Rhesus and Chimp sequences.

Rhesus and Chimp sequences) were given as input sequences for MotifScanner. We used the TRANSFAC Vertebrates motif model and the EPD Vertebrates background model. The prior probability of finding copy of a motif was set to 0.2 and the order of the background model was set to 1. The results were provided in a gff file, which contains the motif id, the motif sequence and the nucleotide position for the identified motifs for each specie. However, we found a huge amount of motifs, many of them may not be real transcription factor binding sites. A kind of filtering in order to find relevant motifs can be done by comparing the results of MotifScanner to the DiAlign regions. If a motif and a DiAlign region have any overlapping nucleotide, it is supposed to be more likely that the motif is a real transcription factor binding site. The overlapping regions were calculated and illustrated by a Perl script. (Figure 9-12) Red regions correspond to the overlapping regions. Blue stretches indicate the motifs and the green ones are the DiAlign similarity regions. For the second ORMDL3 dataset (Human, Rhesus, Chimp) the overlapping regions are collected in Table 2.

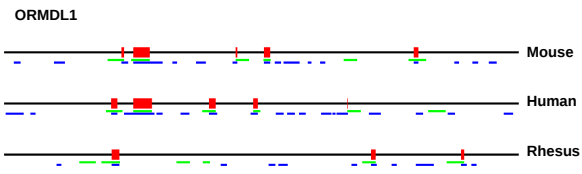


Figure 9: The DiAlign regions (green), found motifs (blue) and the overlapping regions between them (red) for the ORMDL1 nucleotide sequences.

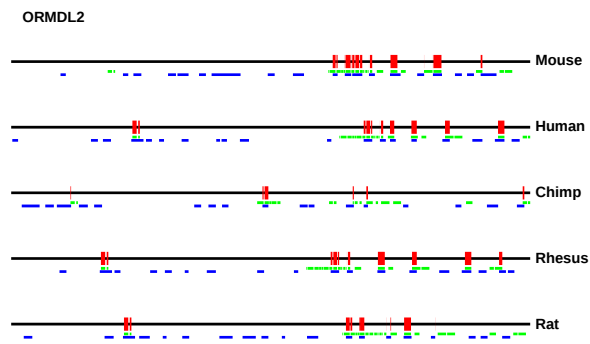


Figure 10: The DiAlign regions (green), found motifs (blue) and the overlapping regions between them (red) for the ORMDL2 nucleotide sequences.

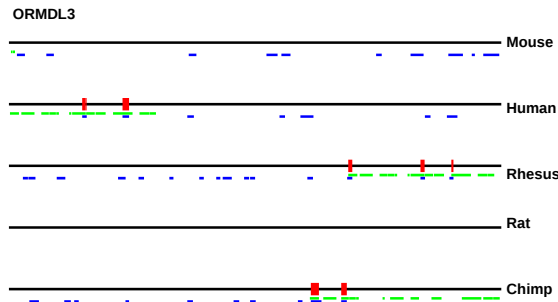


Figure 11: The DiAlign regions (green), found motifs (blue) and the overlapping regions between them (red) for the ORMDL3 nucleotide sequences.

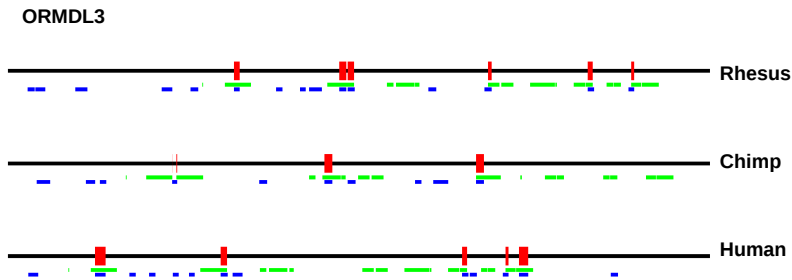


Figure 12: The DiAlign regions (green), found motifs (blue) and the overlapping regions between them (red) for the ORMDL3 reprocessed Human, original Rhesus and Chimp nucleotide sequences.

## 4 Conclusion and future directions

It seems to be a reasonable concept applying all the three methods for a set of orthologous sequences: DiAlign, SAPF and MotifScanner, and considering regions or motifs that were found by all of them. The predictions for the second ORMDL3 dataset unfortunately does not show entire overlap, there were no region found by all the three methods. However, the pairwise comparison of the results determined some regions or motifs that may correspond to

specie	motif id	motif	position
Human	M00033-V\$P300_01	ACCTAGAGTGAGTC	124..137
Human	M00199-V\$AP1_C	GTGAGTCAG	131..139
Human	M00173-V\$AP1_Q2	ACTGACCAAGA	303..313
Human	M00188-V\$AP1_Q4	ACTGACCAAGA	303..313
Human	M00184-V\$MYOD_Q6	CCCACCTGCC	647..656
Human	M00415-V\$AREB6_04	CTGTTTCCT	705..713
Human	M00286-V\$GKLF_01	GAAGGAAAGGTAGG	728..741
Rhesus	M00199-V\$AP1_C	ATGAGTCAG	322..330
Rhesus	M00418-V\$TGIF_01	AGATGTCAGGA	472..482
Rhesus	M00188-V\$AP1_Q4	ACTGACCAAGA	484..494
Rhesus	M00173-V\$AP1_Q2	ACTGACCAAGA	484..494
Rhesus	M00072-V\$CP2_01	CCACAAACCAG	679..689
Rhesus	M00184-V\$MYOD_Q6	CCCACCTGCC	826..835
Rhesus	M00415-V\$AREB6_04	CTGTTTCCT	884..892
Chimp	M00217-V\$USF_C	TCACCTGC	234..241
Chimp	M00001-V\$MYOD_01	AGGCAGGTGCTG	451..462
Chimp	M00277-V\$LMO2COM_01	AGGCAGGTGCTG	451..462
Chimp	M00001-V\$MYOD_01	GGACAGGAGGTG	667..678

Table 2: Found motifs for ORMDL3 reprocessed Human, original Rhesus and Chimp sequences which has got overlap with DiAlign regions.

transcription factor binding sites.

There was correspondence between the SAPF and DiAlign regions in some cases. The Human nucleotides from the position 674 to 694 (ctctgggggtcccaactccagc), the region from 848 to 868 (ctctgggggttctctcaagc) and from 526 to 532 (ggggatg) in Chimp, the nucleotides from 853 to 873 (ctctgggggtcccaactccagc) in Rhesus were found with DiAlign as well. There were motifs found by MotifScanner which overlapped with some regions detected by SAPF. In Rhesus the following overlapping motifs were found: (SAPF: ggggagg, 604-610) M00244-V\$NGFIC\_01 (ATGCGGGGAGG, 599-610) and M00243-V\$EGR1\_01 (ATGCGGGGAGG, 599-610). In Chimp (SAPF: ggggagg, 612-618) M00189-V\$AP2\_Q6 (AGCCCCGGGGAG, 606-617) and M00414-V\$AREB6\_03 (AGGCACCTGGGG, 616-627) had overlap.

The methods could also reveal the changes in some of the orthologous sequences. For example the Chimp ORMDL2 sequence seemed to be very different compared to the other ORMDL2 sequences. FootPrinter and DiAlign revealed these differences. It was also placed to an unexpected place

on the consensus network, created based on MrBayes calculation. DiAlign was able to find the series of shorter high similarity regions, however in some cases it was clear that we may missed to find the corresponding regions in the orthologous upstream sequences. In case of ORMDL3, the first dataset contained Rat and Mouse without significant similarity and Human seemed to be shifted. Therefore, in the second dataset we collected additional nucleotides and ignored 500 nucleotides from the original Human sequence. The role of these 500 nucleotides right before the mRNA could be the subject of further investigation. The Mouse and Rat sequences were completely ignored. MotifScanner tend to find many motifs for each sequence. It is reasonable to assume that some of these motifs are false ones. We tried one way of filtering motifs in order to consider only the relevant ones with comparing them to the results of either SAPF or DiAlign. However, finding relevant motifs can also be an important and interesting challenge. One way for this can be based on calculating the probability that these motifs appear in random sequences as well. Another possible strategy can be searching for series of motifs that appear in the same order in the orthologous sequences.

The overlap rate of the motifs and the DiAlign predicted regions, which indicate the agreement between the results, can be given by the number of overlapping common nucleotides (overlap size, which is the number of nucleotides that are part of a DiAlign region and a motif as well) divided by either the number of nucleotides being in the motifs of the given sequence (motif size) or the number of nucleotides being in DiAlign regions (DiAlign size). In the second set of ORMDL3 sequences the highest overlap rates were found for Human.  $overlap\ size/DiAlign\ size = 0.18$  and  $overlap\ size/motif\ size = 0.35$  which are still rather low. Fortunately for ORMDL1 and ORMDL2 these ratios were higher. If we consider only the overlapping motifs we may perform another filtering with collecting only those motifs that are present in more species.

The second ORMDL3 dataset contained common binding sites (motifs) for Human and Rhesus: M00199-V\$AP1\_C, M00173-V\$AP1\_Q2, M00188-V\$AP1\_Q4, M00184-V\$MYOD\_Q6, M00415-V\$AREB6\_04.

The information about the transcription factor binding sites can be reached in the TRANSFAC database. Additional data about the transcription factor binding sites and the connecting transcription factors would probably give such information, which could help to obtain further results.

It is complicated to take sides with any of the methods because they are based on different concepts. Further investigations may infer the results of one or another method. However, it must not be ignored that the results can highly depend on the given sequences. Therefore, it is always an important question if we found the right piece of our upstream region. If we have chosen

a particular set of sequences it is also an important point to analyze them at first to be able to find the best method for the investigation.

## 5 Acknowledgement

I would like to thank Jotun Hein and István Miklós that I had the opportunity to work on the project and for supervising the work. Thank István Miklós for guiding the project with useful discussions and continuous support. I also wish to thank Ádám Novák, Rahul Satija and Ferenc Huszár for their help and Illés Farkas for his support.

## References

- S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, March 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg268. URL <http://dx.doi.org/10.1093/nar/gkg268>.
- M. Blanchette and M. Tompa. Footprinter: a program designed for phylogenetic footprinting. *Nucleic Acids Research*, 31(13):3840–3842, 2003.
- R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500, July 2003. ISSN 1362-4962. URL <http://nar.oxfordjournals.org/cgi/content/abstract/31/13/3497>.
- L. Hjelmqvist, M. Tuson, G. Marfany, E. Herrero, S. Balcells, and R. Gonzalez-Duarte. Ormdl proteins are a conserved new family of endoplasmic reticulum membrane proteins. *Genome Biology*, 3:research00, 2002. URL <http://www.citebase.org/abstract?id=oai:biomedcentral.com:gb-2002-3-6-research0027>.
- D. H. Huson and D. Bryant. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol*, 23(2):254–267, 2006. doi: 10.1093/molbev/msj030. URL <http://mbe.oxfordjournals.org/cgi/content/abstract/23/2/254>.
- V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. *Transfac(r)*

- and its module transcompel(r): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue), January 2006. ISSN 1362-4962. URL <http://view.ncbi.nlm.nih.gov/pubmed/16381825>.
- M. F. Moffatt, M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S. A. G. Willis-Owen, K. C. C. Wong, T. Illig, C. Vogelberg, S. K. Weiland, E. von Mutius, G. R. Abecasis, M. Farrall, I. G. Gut, M. G. Lathrop, and W. O. C. Cookson. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, 448:470–473, July 2007. ISSN 0028-0836. doi: 10.1038/nature06014. URL <http://dx.doi.org/10.1038/nature06014>.
- B. Morgenstern. DiAlign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, 1999.
- R. Satija, L. Pachter, and J. Hein. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics*, 24(10):1236–1242, 2008.
- E. Wingender, P. Dietze, H. Karas, and R. Knüppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic Acids Res*, 24(1):238–241, January 1996. ISSN 0305-1048. URL <http://view.ncbi.nlm.nih.gov/pubmed/8594589>.