Methodology to identify conserved sequence blocks in an alignment of intergenic sequences

Pieter Monsieurs, Joke Allemeersch, Kathleen Marchal, Bart De Moor
ESAT-SISTA/COSIC, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee Belgium
Email: {Pieter.Monsieurs,Kathleen.Marchal}@esat.kuleuven.ac.be

Introduction

Genes controlled by a specific regulatory protein can be found by a genome wide screening of the intergenic regions for the presence of a corresponding regulatory binding site (i.e. regulatory motif). Performing such an *in silico* motif detection will frequently result in a high number of false positives (i.e. genes displaying a motif hit without biological relevance). This problem is exacerbated when the length of the binding site of the regulatory motif is small. Checking whether the motif is conserved in the promoter region of close homologs, can help reducing the number of false positives. The fact that evolution tends to conserve the motif in close homologs, points towards the biological validity of the motif. Cross species comparison of the conservation of a motif can be done by e.g. checking whether the motif is present in the intergenic sequences of the orthologs of the original sequence (by a pattern search). Also here the detection of a particular motif in all orthologs is no guarantee. The detected hits in other genomes might also be spurious. The problem is how to validate whether in a set of orthologous sequences the detected motifs correspond to different false positive hits or whether they truly reflect conserved motifs.

Earlier observations showed us that there is a high similarity between intergenic regions of close homologs of gamma proteobacteria (analysis performed on *Salmonella enterica*, *Escherichia coli*, *Shigella flexneri* and *Yersinia pestis*). This implicates that when local alignments are created based on the positions of regulatory motifs in close homologs, not only the motif is conserved, but also the context in which the motif is located. On the contrary, when an alignment based on the positions of the regulatory motif doesn't display this context similarity, the motif hit usually can be classified as a false positive (Fig.1). We developed an approach that exploits this context similarity in intergenic regions to make a distinction between false positive motif hits in the intergenic regions of the orthologs and truly conserved motifs. The positions of motif hits in the intergenic sequences of the orthologs were used as alignment initialisation seeds. The degree of conservation in the local neighbourhood of the aligned motifs was evaluated by comparing the observed degree of conservation to the degree of conservation that can be expected in a random dataset. Based on this statistical evaluation a certain confidence to the predictions could be assigned.

Results

Detection of the longest conserved block

As mentioned before multiple local alignments were generated using detected motifs as seeds. In such local alignments we searched for short conserved DNA-sequences (i.e. a conserved block) in the neighbourhood of the putative motif i.e. 50 bp in the upstream and downstream region of the motif. This allowed deciding whether the context of the motif is conserved. Conserved blocks were identified based on the consensus score of the respective blocks.

$$consensus - score = 2 - \frac{1}{W}\sum_{l=1}^{W}\sum_{b(actg)} q_l^b . \log q_l^b$$

The reason that the consensus score was chosen as a measurement of this conservation of intergenic regions, is to prevent that a generally well conserved block but with one or a few less conserved positions is interpreted as two different short conserved blocks. The latter would be the case if we used the conservation at each individual position in the alignment as a selection criterion. Starting at the first column of the context of the motif (i.e. the aligned nucleotides 50 bp upstream of the start of the regulatory motif), the consensus score was calculated for that position. As long as this consensus score was above a user defined threshold, the block was extended with an additional column from the alignment and the consensus score was recalculated. When a position is reached where the consensus score fell below the threshold score, this position was regarded as the end of the conserved block. The analysis started all over again searching a new conserved block beginning from the second column in the alignment i.e. each column in the alignment was regarded as a potential start position of a conserved block.

Generating random sets

To get a more quantitative idea about the likelihood of observing the largest conserved block, the length of this block in the real biological alignment was evaluated by comparing it to the distribution of the lengths of conserved blocks that can be identified in random sets. The random sets contained sequences derived from species corresponding to the combination of species present in the real biological datasets. Each random dataset consisted of at least 1000 alignments of intergenic regions. The datasets were generated in such a way that the observed high pairwise sequence homology between intergenic sequences derived from similar species (e.g. high similarity in intergenic regions between *Salmonella typhimurium* and *Salmonella typhi*, between different strains of *Escherichia coli*, …) was taken into account, but not the observed overall similarity between the intergenics of different species (e.g. between *Escherichia coli* and *Salmonella*). Sequences derived e.g. from *E. coli* are not independent samples. Their high mutual similarity increases the probability of finding a conserved block with a certain length in a dataset consisting of three orthologous *E. coli* genes as compared to the probability of finding it in an alignment of three independent intergenic sequences of *E. coli*. Not taking into account this intraspecies homology would result in an overestimation of the number of truly conserved blocks.

Generating the distributions for the random datasets

On these random datasets, a *de novo* motif finding algorithm can be performed. The positions of the detected motifs can be used as initialisation to construct an alignment of intergenic regions. On alignments generated as such from random sets, we identified the different conserved blocks by counting in each random set the number of motifs of length w (1, 2, 3,

4,…) respectively by using a sliding window of length w. Using the sliding window each position in the sequence can be a potential start of the motif. Because the counts for a block of length w include the counts of a block w-1, the empirically derived distribution corresponds to a cumulative distribution.

As such a histogram was built that displays for each possible length of the conserved block the number of its occurrences in the random set. The parametric probability distribution that best fits these data was the logarithmic series distribution. This parametric distribution was used to derive a cumulative probability distribution for each length of a conserved block. The p-value derived from this distribution displays the chance to observe in the random data set a conserved block with length equal or higher than the specified length.

As such, for each value of the longest block in the true data set, a corresponding p-value can be calculated. Alignments with a high p-value are rejected because they are assumed to be false positives.

Figure Legends

Figure 1:

Examples of local alignments based on the position of the PhoP regulatory motif. A. Example of a local alignment of the *phoP*-gene itself (p-value 0.012) where information from the regulatory motif is used as initialisation for the alignment. B. An example of a biological relevant alignment in the *ybbU/allR*-gene (p-value 0.00001). Several conserved blocks can be detected in this alignment. C. Alignment in the *tdcA*-gene (p-value 0.18). In this case, only short conserved blocks can be detected. This is probably not the correct alignment of the intergenic region, and as a consequence, the alignment will be considered as false positive.

Figure 2:

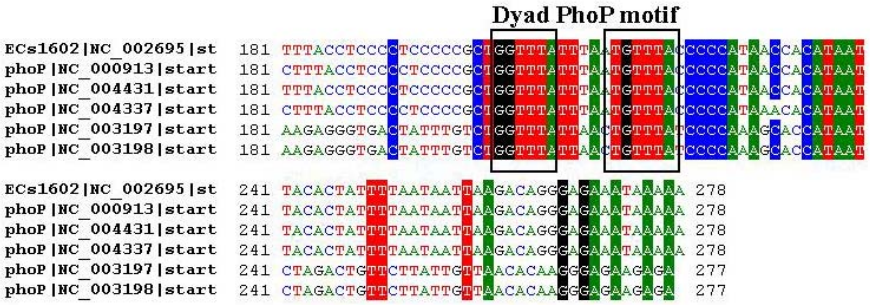
Distribution of the length of conserved DNA-regions in a random dataset containing members from four species (*Salmonella enterica* (2 serovars)*, Shigella flexneri, Escherichia coli* (3 strains) and *Yersinia pestis* (2 strains). The histogram displays the distribution of the normalised number of conserved blocks with respect to their length. The full line gives the parametric probability distribution that best fits these data i.e. the logarithmic series distribution.

Figure 3:

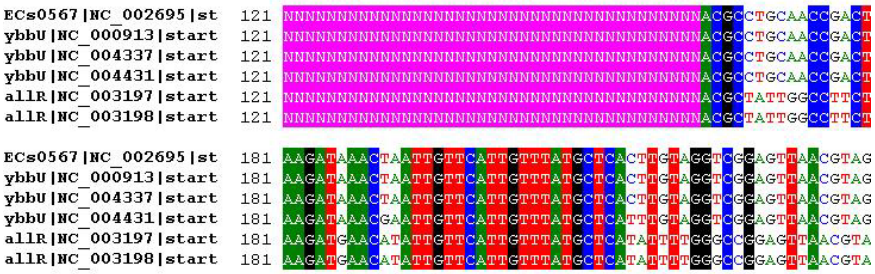
The progress of the consensus score over the alignment of intergenic regions derived from PhoP orthologs. In the X-axis the position in the alignment is displayed relative to the detected motif, while in the Y-axis the consensus score is displayed. Only when the consensus score exceeds the user-defined threshold, the score is plotted.

Figure 1

A.


Dyad PhoP motif
ECs1602|NC_002695|st 181 TTTACCTCCCCTCCCCCGCTGGTTTATTTAATGTTTACCCCCATAACCACATAAT
phoP|NC_000913|start 181 CTTTACCTCCCCTCCCCGCTGGTTTATTTAATGTTTACCCCCATAACCACATAAT
phoP|NC_004431|start 181 TTTACCTCCCCTCCCCCGCTGGTTTATTTAATGTTTACCCCCATAACCACATAAT
phoP|NC_004337|start 181 CTTTACCTCCCCTCCCCGCTGGTTTATTTAATGTTTACCCCCATAAACACATAAT
phoP|NC_003197|start 181 AAGAGGGTGACTATTTGTCTGGTTTATTAACTGTTTATCCCCAAAGCACCATAAT
phoP|NC_003198|start 181 AAGAGGGTGACTATTTGTCTGGTTTATTAACTGTTTATCCCCAAAGCACCATAAT
ECs1602|NC_002695|st 241 TACACTATTTTAATAATTAAGACAGGGAGAAATAAAAA 278
phoP|NC_000913|start 241 TACACTATTTTAATAATTAAGACAGGGAGAAATAAAAA 278
phoP|NC_004431|start 241 TACACTATTTTAATAATTAAGACAGGGAGAAATAAAAA 278
phoP|NC_004337|start 241 TACACTATTTTAATAATTAAGACAGGGAGAAATAAAAA 278
phoP|NC_003197|start 241 CTAGACTGTTCTTATTGTTAACACAAGGGAGAAGAGA 277
phoP|NC_003198|start 241 CTAGACTGTTCTTATTGTTAACACAAGGGAGAAGAGA 277

B.


ECs0567|NC_002695|st 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCCTGCAACCGACT
ybbU|NC_000913|start 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCCTGCAACCGACT
ybbU|NC_004337|start 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCCTGCAACCGACT
ybbU|NC_004431|start 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCCTGCAACCGACT
allR|NC_003197|start 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCTATTGGCCTTCT
allR|NC_003198|start 121 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACGCTATTGGCCTTCT
ECs0567|NC_002695|st 181 AAGATAAACTAATTGTTCATTGTTTATGCTCACTTGTAGGTCGGAGTTAACGTAG
ybbU|NC_000913|start 181 AAGATAAACTAATTGTTCATTGTTTATGCTCACTTGTAGGTCGGAGTTAACGTAG
ybbU|NC_004337|start 181 AAGATAAACTAATTGTTCATTGTTTATGCTCACTTGTAGGTCGGAGTTAACGTAG
ybbU|NC_004431|start 181 AAGATAAACGAATTGTTCATTGTTTATGCTCATTTGTAGGTCGGAGTTAACGTAG
allR|NC_003197|start 181 AAGATGAACATATTGTTCATTGTTTATGCTCATATTTTGGGCCGGAGTTAACGTA
allR|NC_003198|start 181 AAGATGAACATATTGTTCATTGTTTATGCTCATATTTTGGGCCGGAGTTAACGTA

C.

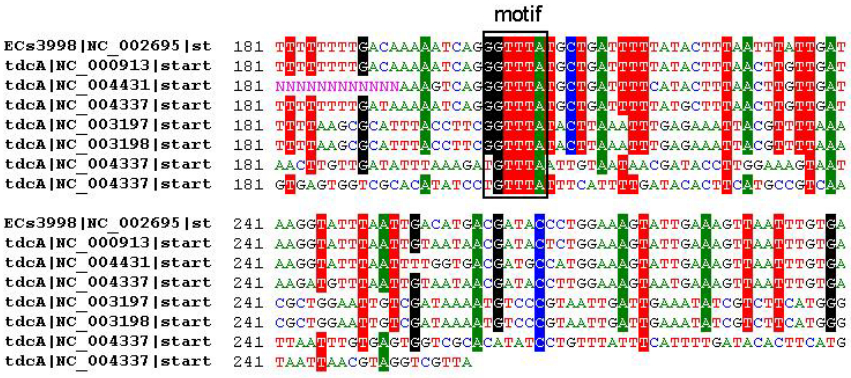

motif
ECs3998|NC_002695|st 181 TTTTTTTTGACAAAAATCAGGGTTTATGCTGATTTTTATACTTTAATTTATTCAT
tdcA|NC_000913|start 181 TTTTTTTTGACAAAAATCAGGGTTTATGCTGATTTTTATACTTTAACTTGTTGAT
tdcA|NC_004431|start 181 NNNNNNNNNNNNAAAGTCAGGGTTTATGCTGATTTTCATACTTTAACTTGTTGAT
tdcA|NC_004337|start 181 TTTTTTTTTGATAAAAATCAGGGTTTATGCTGATTTTTATGCTTTAACTTGTTGAT
tdcA|NC_003197|start 181 TTTTAAGCGCATTTACCTTCGGTTTATACTTAAATTTGAGAAATTACGTTTTAAA
tdcA|NC_003198|start 181 TTTTAAGCGCATTTACCTTCGGTTTATACTTAAATTTGAGAAATTACGTTTTAAA
tdcA|NC_004337|start 181 AACTTGTTGATATTTAAAGATGTTTATTGTAATAACGATACCTTGGAAAGTAAT
tdcA|NC_004337|start 181 GTGAGTGGTCGCACATATCCTGTTTATTTCATTTTGATACACTTCATGCCGTCAA
ECs3998|NC_002695|st 241 AAGGTATTTAATTGACATGACGATACCTGGAAAGTATTGAAAGTTAATTTGTGA
tdcA|NC_000913|start 241 AAGGTATTTAATTGTAATAACGATACTCTGGAAAGTATTGAAAGTTAATTTGTGA
tdcA|NC_004431|start 241 AAGGTATTTAATTTTGGTGACGATGCATGGAAAGTATTGAAAGTTAATTTGTGA
tdcA|NC_004337|start 241 AAGATGTTTAATTGTAATAACGATACTTGGAAAGTAATGAAAGTTAATTTGTGA
tdcA|NC_003197|start 241 CGCTGGAATTGTCGATAAAATGTCCCGTAATTGATTGAAATATCGTCTTCATGGG
tdcA|NC_003198|start 241 CGCTGGAATTGTCGATAAAATGTCCCGTAATTGATTGAAATATCGTCTTCATGGG
tdcA|NC_004337|start 241 TTAATTTGTGAGTGGTCGCACATATCCTGTTTATTTCATTTTGATACACTTCATG
tdcA|NC_004337|start 241 TAATTAACGTAGGTCGTTA

Figure 2

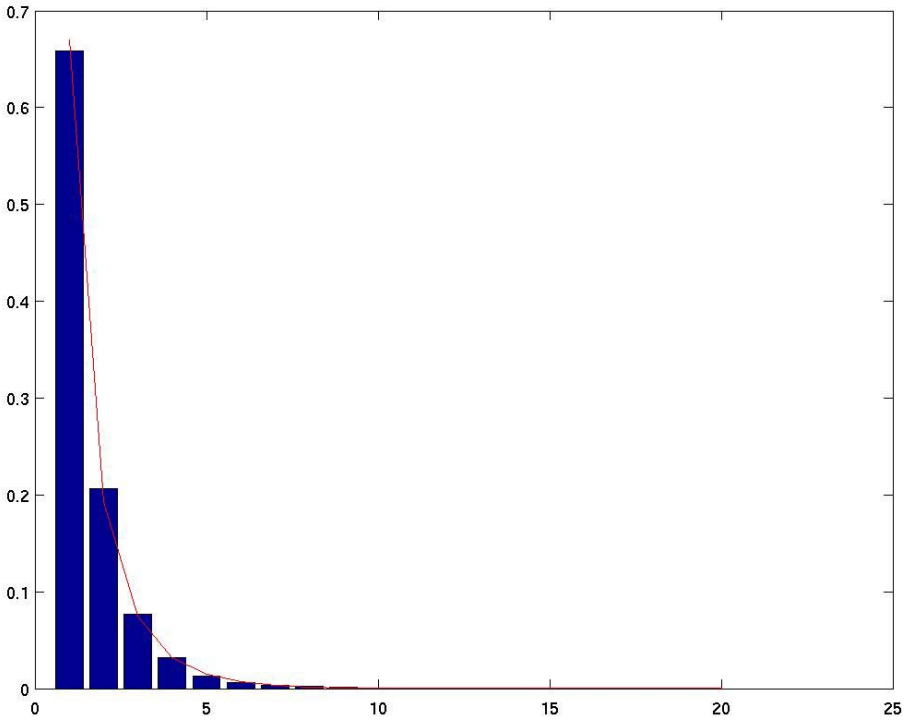
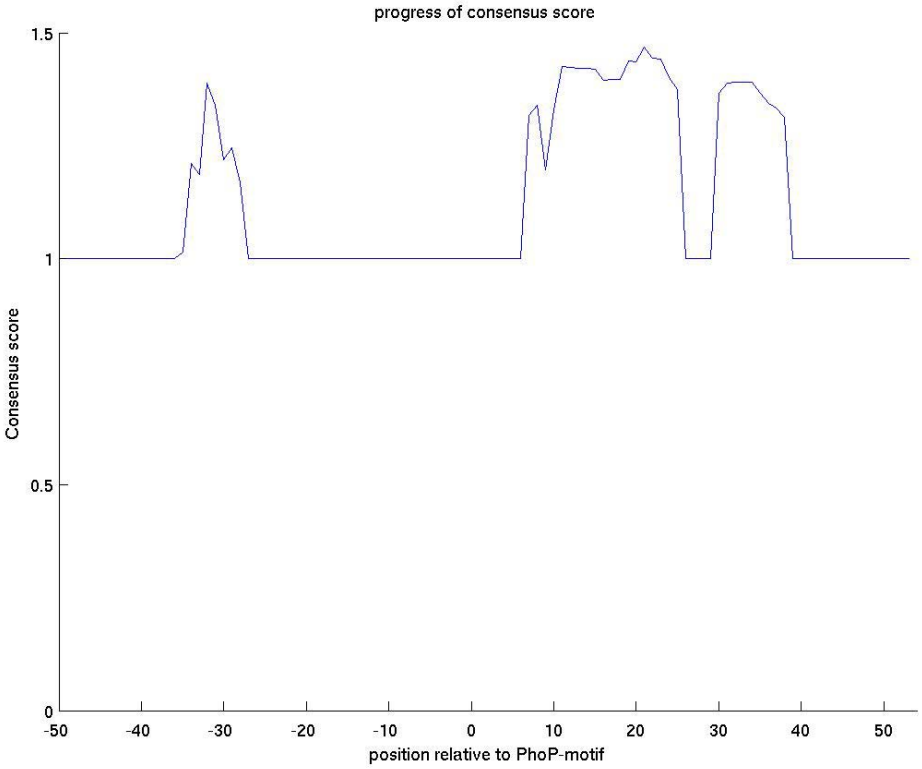
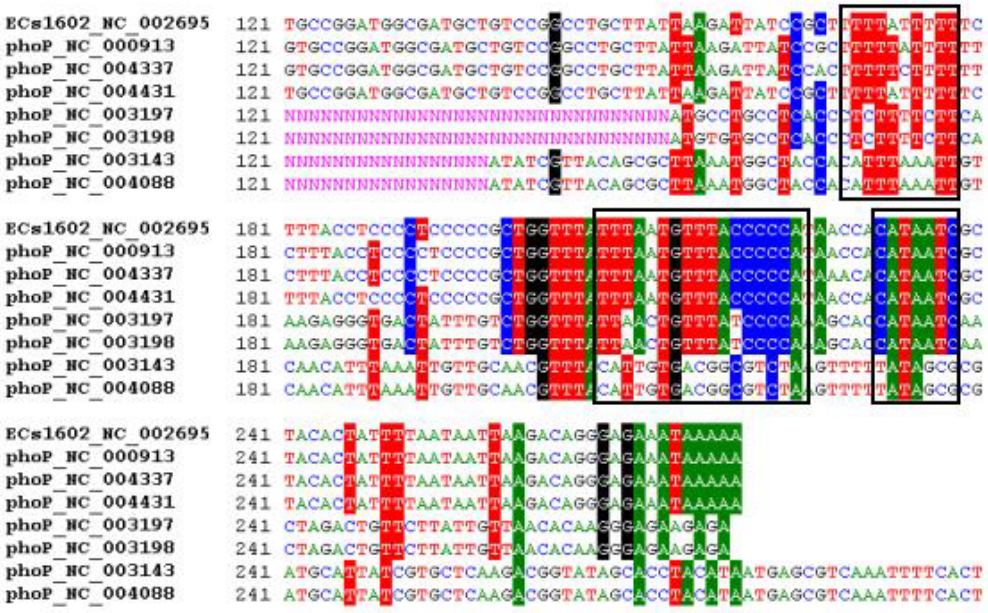


Figure 3


progress of consensus score
Consensus score
1.5
1
0.5
0
-50
-40
-30
-20
-10
0
10
20
30
40
50
position relative to PhoP-motif


ECs1602_NC_002695 121
phoP_NC_000913 121
phoP_NC_004337 121
phoP_NC_004431 121
phoP_NC_003197 121
phoP_NC_003198 121
phoP_NC_003143 121
phoP_NC_004088 121
ECs1602_NC_002695 181
phoP_NC_000913 181
phoP_NC_004337 181
phoP_NC_004431 181
phoP_NC_003197 181
phoP_NC_003198 181
phoP_NC_003143 181
phoP_NC_004088 181
ECs1602_NC_002695 241
phoP_NC_000913 241
phoP_NC_004337 241
phoP_NC_004431 241
phoP_NC_003197 241
phoP_NC_003198 241
phoP_NC_003143 241
phoP_NC_004088 241