Development of a heuristic for the selection of alignment seeds from a list of motifs generated by Gibbs sampling

K. Marchal*,Pieter Monsieurs*, Karen Lemmens *, B. De Moor*

* ESAT-SCD, Dept. Elect. Eng., Catholic University of Leuven,

Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

{*kathleen.marchal,bart.demoor*}*@kuleuven.ac.be*

Abstract

In this technical report we describe a heuristic that can be used to select motifs from a list of motifs generated by Motif sampler that will most likely generate biological relevant alignments. The heuristic was found by comparing motif scores obtained from a random set with those obtained in a genuine dataset by Gibbs sampling.

1 Generation of random sets

For each dataset a corresponding random set was generated. To each sequence in the genuine dataset corresponds in the random set a sequence selected randomly from the respective organism and trimmed to the same length as the length of the original sequence in the genuine dataset. Because we aimed at constructing local alignments, motifs that reflect the correlation between data are of primary interest and do not need to be filtered out. Therefore, we explicitly did not model the observed sequence similarity between intergenic regions of evolutionary related organisms (e.g. 70 % similarity observed for *S. typhimurium* and *E. coli*) in the random sets. For each dataset, 5 random simulated datasets were constructed and the same Gibbs sampling strategy was performed as with the genuine dataset. For each dataset, the LL-scores for all 5 simulations were averaged and used to determine the heuristic described below.

2 Heuristic for selecting potential seeds for local multiple alignments

The heuristic to select motifs that could serve as seeds for multiple local alignments of the intergenic regions was developed by comparing the LL-score, consensus score and the number of instances of the motifs obtained within the genuine dataset, against the values of these scores obtained from motifs retrieved in the corresponding random datasets. Fig. 1 A. compares the distribution of the LL-score of all motifs from the genuine set with those of the random set (normalized by the number of simulations) for a few representative examples. From these results, it could be concluded that on average, the genuine dataset displays motifs with a higher LL-score (distribution moved towards the right) than the random set. This indicates that compared to the random dataset, the genuine dataset contains correlated intergenic sequences, and thus potential seeds for biologically relevant multiple local alignments. However, as appears

from Fig. 1 A., the distinction in the distribution between the random and the genuine dataset is not always as clearcut. For pmrC for instance, a considerable overlap between the distribution of the LL-scores obtained from the random and the genuine sets, is observed. This overlap in distributions tends to become larger as the overall similarity between the intergenic sequences in the genuine dataset becomes lower (less correlation present in the data), and the degree of correlation in the genuine dataset starts resembling this of the random dataset. This indicates that the stringency of using the results obtained from the random set, in selecting a single critical value (e.g. 95 percentile) on the LL-score, depends on the overlap between the distribution of the genuine and the random sets. Summarizing, among the high scoring motifs in the genuine dataset, a considerable number of motifs is present that might not result in biological relevant local alignments (false positives). Therefore, we relied, in addition to the LL-score, on additional selection criteria. In Fig. 1 B is illustrated for the representative examples, how the consensus score relates to the number of instances and the LL-score in the genuine set. The LL-score, by definition, increases as the motif becomes more conserved (higher consensus score) and as it is more statistically overrepresented in the dataset (higher number of instances) [1] [2]. Because we were interested in producing local alignments, we expected motifs that can be used as seeds for such alignments to be highly conserved (high consensus score) and to occur preferentially once in each sequence of the genuine dataset. The regions where the number of motif instances was equal to the number of sequences in the genuine dataset are indicated (by boxes). Usually as more instances are present, the consensus score decreases as the motif becomes more degenerated (for instance in example of pmrC). Occasionally the LL-score, which calculates a trade off between the number of instances and the consensus score can still increase as more instances are present. However, local alignments obtained from such seeds were less obvious (see supplementary information).

3 Conclusion

Taken together, these observations resulted in the following heuristic: based on the distribution of the LL-score in the random set, we set a critical threshold on the minimal LL-score. Motifs with a score above this threshold that occurred preferentially once in each sequence were used (starting with those motifs that had the highest consensus score) until a reliable alignment was obtained. Moreover, we preferentially selected motifs that occurred in the maximal number of species. As such, the motifs within the region indicated in the figures all consist of potential seeds for biological relevant alignments.

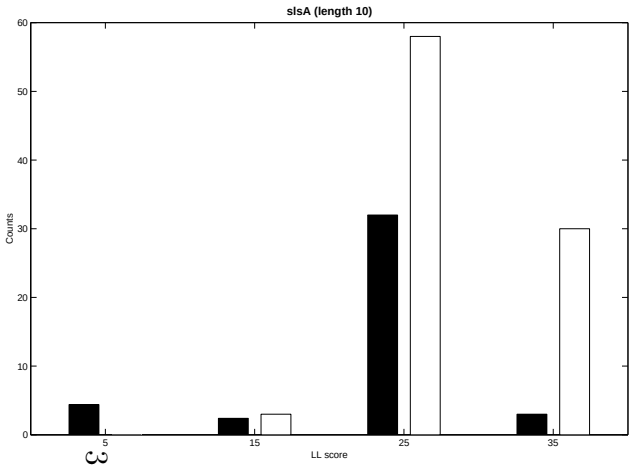
4 Charcteristics of alignements obtained from different seeds

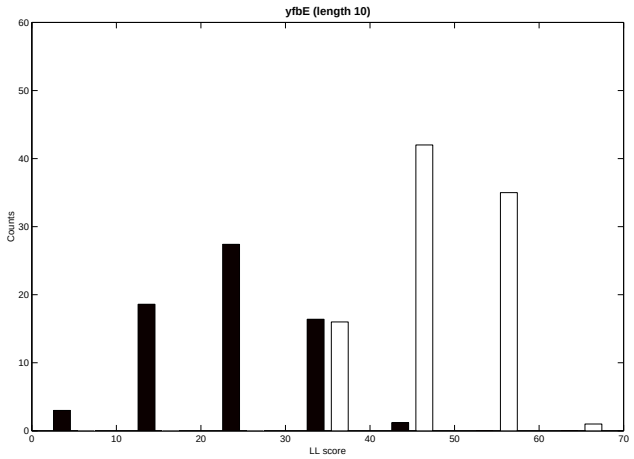
In the figures 2 3 4 the alignments can be judged based on a visual inspection. Because of the observed high overall homology between the intergenic sequences of close homologs derived from *E. coli, Shigella, Salmonella* and *Yersinia* species, we expect that in general in a good local multiple alignment the pairwise homology reflects the phylogenetic distance between the species (degree of conservation decreases gradually with the phylogenetic distance).

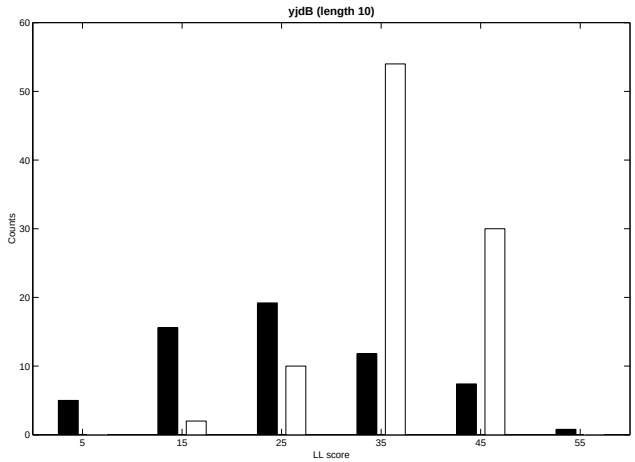
An example is given below:

Three different alignments obtained for *yjdB*. Alignment A Fig. 2 was obtained by using

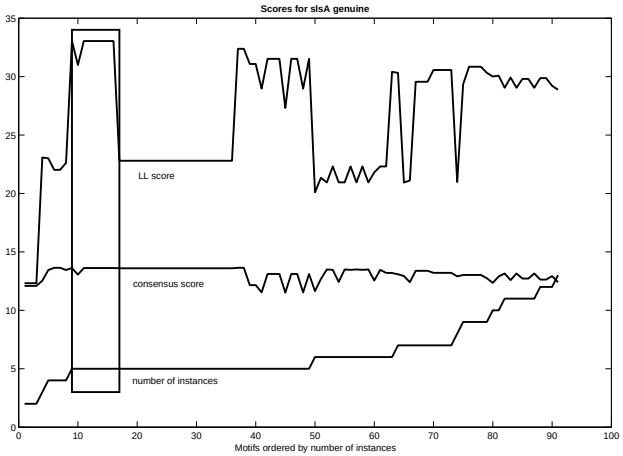
A.

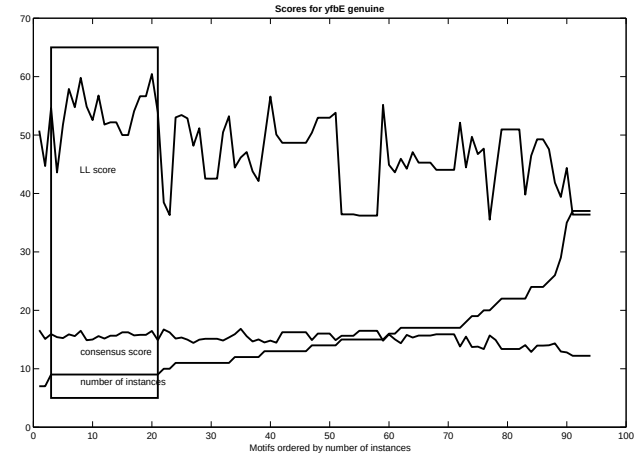

slsA (length 10)
Counts
LL score

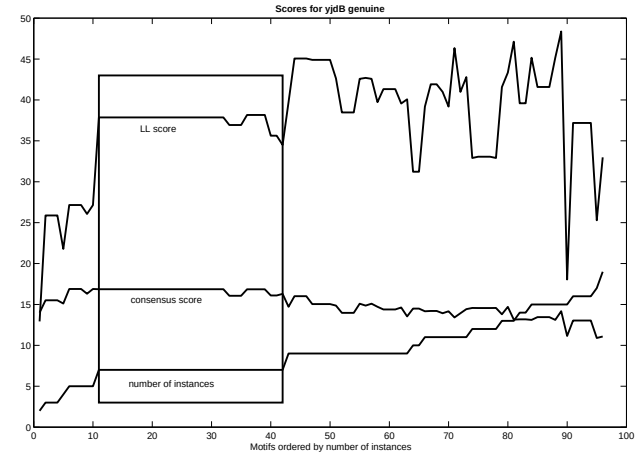

yfbE (length 10)
Counts
LL score


yjdB (length 10)
Counts
LL score

B.


Scores for slsA genuine
LL score
consensus score
number of instances
Motifs ordered by number of instances


Scores for yfbE genuine
LL score
consensus score
number of instances
Motifs ordered by number of instances


Scores for yjdB genuine
LL score
consensus score
number of instances
Motifs ordered by number of instances

a motif (length 10) with high LL-score (48.3) as seed but of which the number of instances (15) exceeded the number of sequences (10). The *E. coli* intergenics are aligned but for the *Salmonella* intergenics the alignment is less clear (bad alignment, not reflecting the phylogenetic relationship).

Alignment B Fig. 3 was obtained by using a motif (length 10) with a lower LL-score (45) as seed of which the number of instances (9) was equal to the number of sequences (9). The dataset contained 3 close homologs in *E. coli* O157 in which the motif was detected. However, the motif was not really conserved (the consensus score was low) decreasing the confidence in this alignment.

Alignment C Fig. 4 was obtained by using a motif (length 10) with a lower LL-score (38) as seed of which the number of instances (7) was equal to the number of sequences (7). These 7 sequences were true orthologs. In this case the highest consensus score was obtained. As compared to the alignment A, here the degree of conservation between the sequences in this alignment reflects the phylogenetic relationship between *E. coli* and *Shigella* and between *E. coli* and *Salmonella*. This alignment was retained.

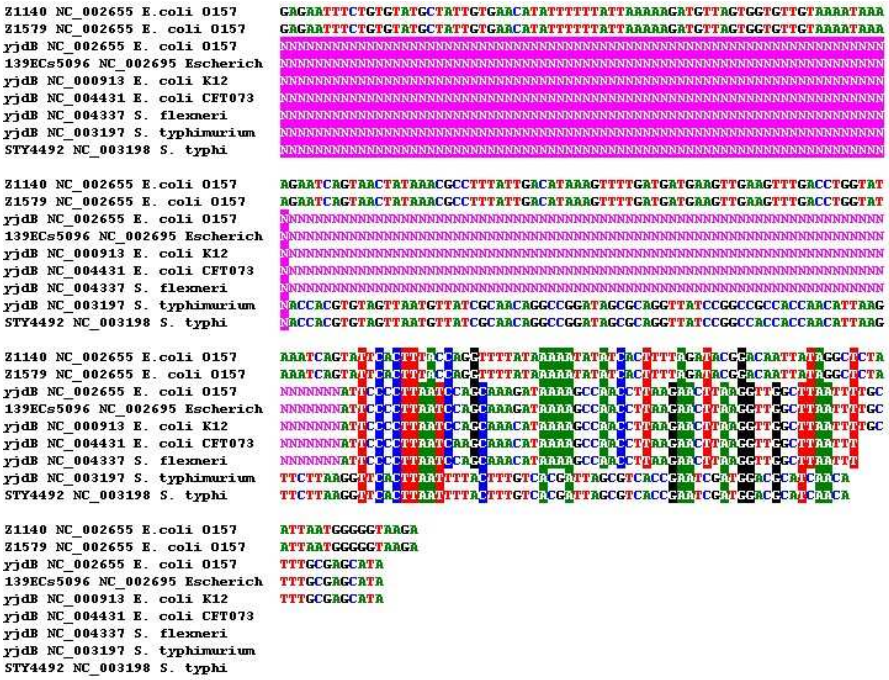


Figure 2: Example of alignments of *yjdB* obtained from 3 different seeds: alignment A

References

[1] Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput. Biol.* **9** (2002) pp. 447-464.

[2] Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J: Genome-specific higher-order background models to improve motif detection. *Trends Micr.* **11** (2002) pp. 61-66.

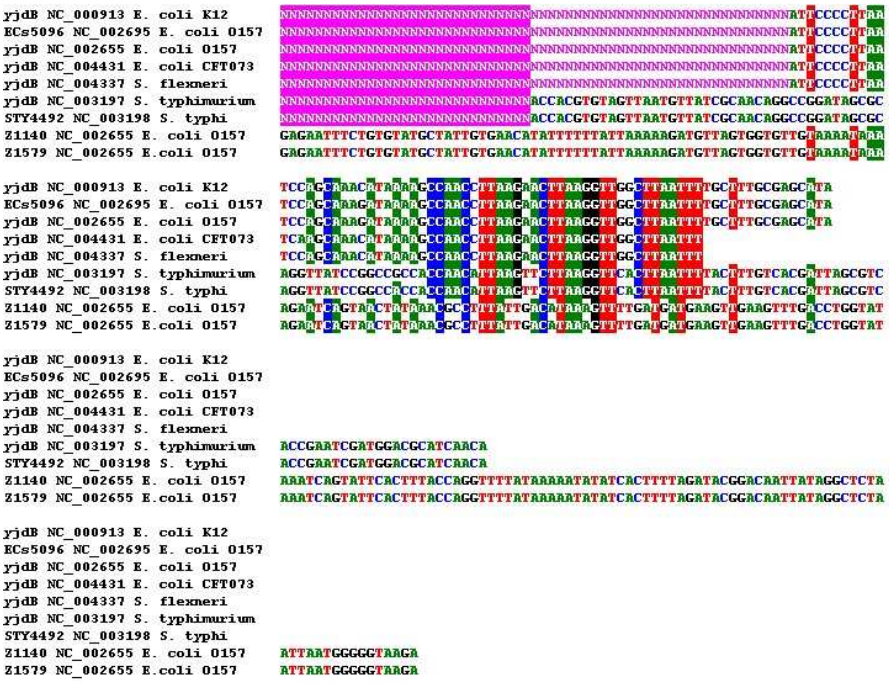


Figure 3: Example of alignments of *yjdB* obtained from 3 different seeds: alignment B

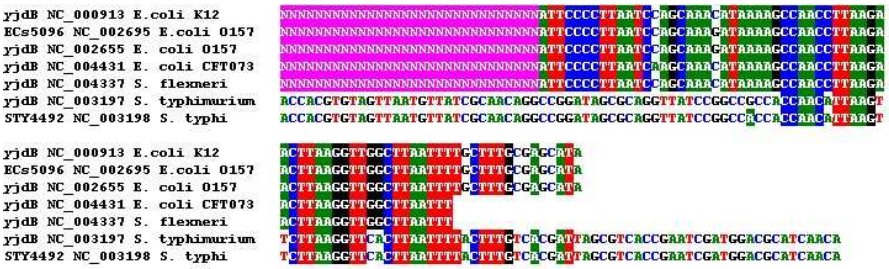


Figure 4: Example of alignments of *yjdB* obtained from 3 different seeds: alignment C