An Alignment-free Method for Detection of Functional Conservation of Regulatory Sequences

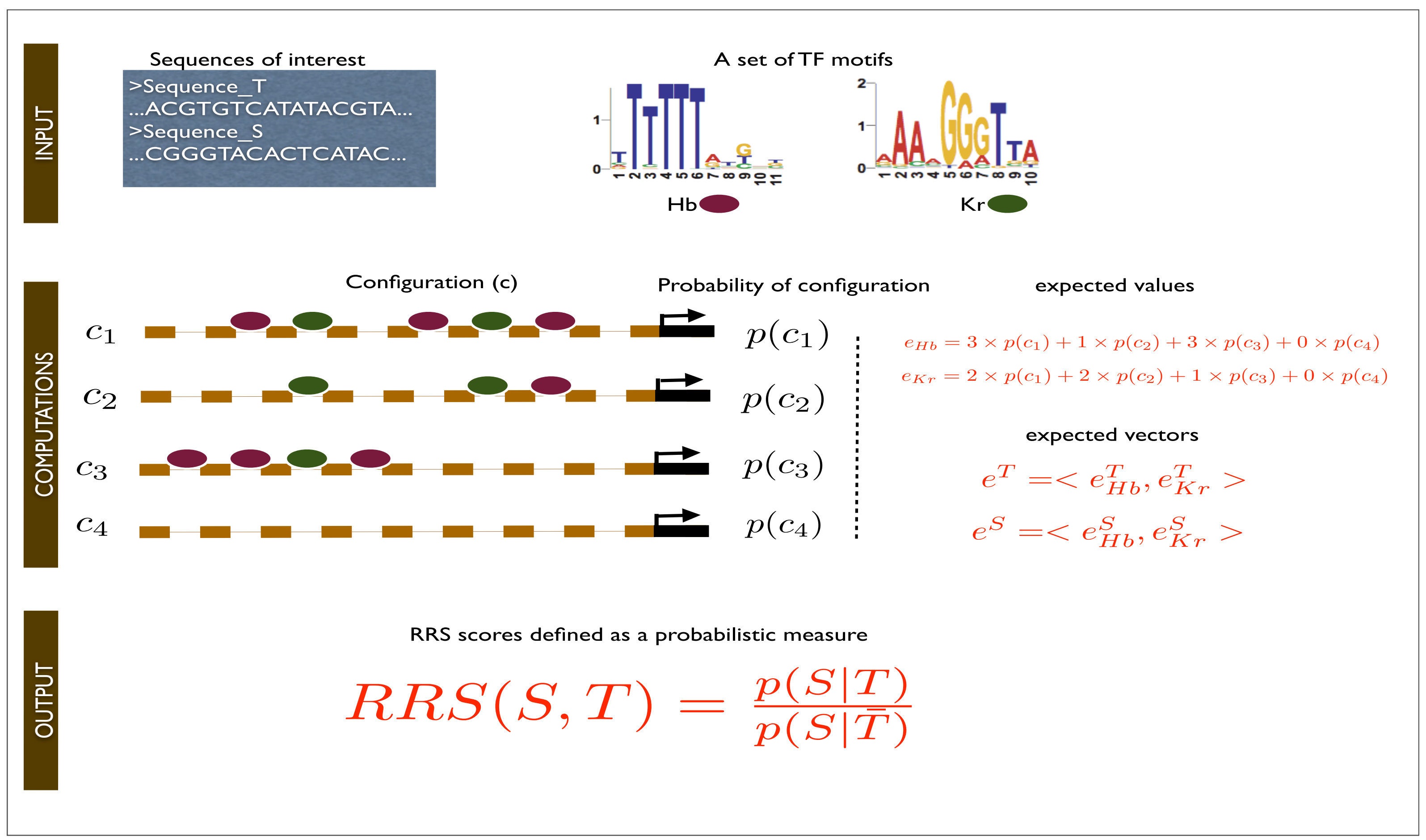
Hashem Koohy, Sascha Ott and Georgy Koentges
Systems Biology Centre and Biological Sciences,
University of Warwick, UK.
Hashem.Koohy@warwick.ac.uk

1 Introduction

Cis-regulatory modules (CRMs) can drive precise spatio-temporal gene expression patterns. Some recent studies show that CRMs may function similarly despite substantial sequence divergence. This implies that firstly alignment-based sequence comparison tools are not applicable for further decoding such CRMs and secondly that some of the CRMs must share common patterns that drive almost identical regulatory outputs but possibly with different arrangements of binding sites. Here, we present our Regulatory Region Scoring (RRS) method which is based on potential distribution of transcription factors. Comparing two regulatory sequences for (dis)similarity has important applications such as:

1. Detect functionally conserved enhancer regions in orthologous genomes even if the enhancer regions do not align.
2. For a given enhancer, detect other enhancers in the same genome that are likely to have a similar function.

2 Modelling



INPUT

Sequences of interest

>Sequence_T
...ACGTGTCATATACGTA...
>Sequence_S
...CGGGTACACTCATAC...

A set of TF motifs

COMPUTATIONS

Configuration (c)

Probability of configuration

expected values

c_1 $p(c_1)$

c_2 $p(c_2)$

c_3 $p(c_3)$

c_4 $p(c_4)$

$e_{Hb} = 3 \times p(c_1) + 1 \times p(c_2) + 3 \times p(c_3) + 0 \times p(c_4)$

$e_{Kr} = 2 \times p(c_1) + 2 \times p(c_2) + 1 \times p(c_3) + 0 \times p(c_4)$

expected vectors

$e^T = < e^T_{Hb}, e^T_{Kr} >$

$e^S = < e^S_{Hb}, e^S_{Kr} >$

OUTPUT

RRS scores defined as a probabilistic measure

$$RRS(S,T) = \frac{p(S|T)}{p(S|\bar{T})}$$

Wednesday, 17 March 2010

Figure 1: An schematic illustration of the RRS

Our model is built of two components that are briefly outline as: :

1. *First component of the RRS: Definition of e-vectors:*
   - For each possible configuration c, obtain the weight of configuration.
   - Obtain the probability of each valid configuration
     $$P(c) = \frac{W(c)}{\sum_{c \in C} W(c)} \tag{1}$$
   - For each transcription factor M_i, compute the expected number of occurrences of M_i in the sequence T i.e.,
     $$e^T_{M_i} = \sum_{c \in VC} P(c) I_{M_i}(c)$$
     and then for the given sequence and set of TFs, define e-vector as :
     $$e^T = < e^T_{M_1}, \cdots, e^T_{M_n} >$$
     We may assume that these e-vetors are normalized to make them independent from the length of sequences.
2. *Second Component of the RRS: Comparison of e-vectors*
   Having some prior knowledge from our template sequence T, for each of the test sequences S we test if it is more likely to come from the same family as the template sequence rather than coming from the genomic background:
   $$\frac{P(E(S)|T)}{P(E(S)|\bar{T})} = \prod_{i=1}^{i=n} \frac{pnorm(s_i|\mu = t_i, \sigma = \sigma_R)}{pnorm(s_i|\mu = \mu_R, \sigma = \sigma_R)}$$

3 Method Comparison

Prediction of functional conservation of non-alignable enhancers from D.melanogaster

Data Set	#CRMs	Top$K$(Random)	D2Z	RRS
EYES	17	136(68)	79(58%)	92(67%)
PNS	23	253(127)	156(62%)	152(60%)
BLASTODERM	82	300(150)	220(73%)	197(65%)
TRACHEAL	9	36(18)	27(75%)	22(61%)

Table 1: Comparison of performance of the RSS vs D2Z over same sets of data. The first column shows four sets of fly known enhancers from REDfly database. The second column is the number of enhancers within each of the sets, the third shows the number of top scoring pairs and the expected number of predictions by chance. The fourth and fifth columns are number of correct predictions (and percentage of correct predictions) from each set of enhancers from D2Z and RRS respectively.

4 Results

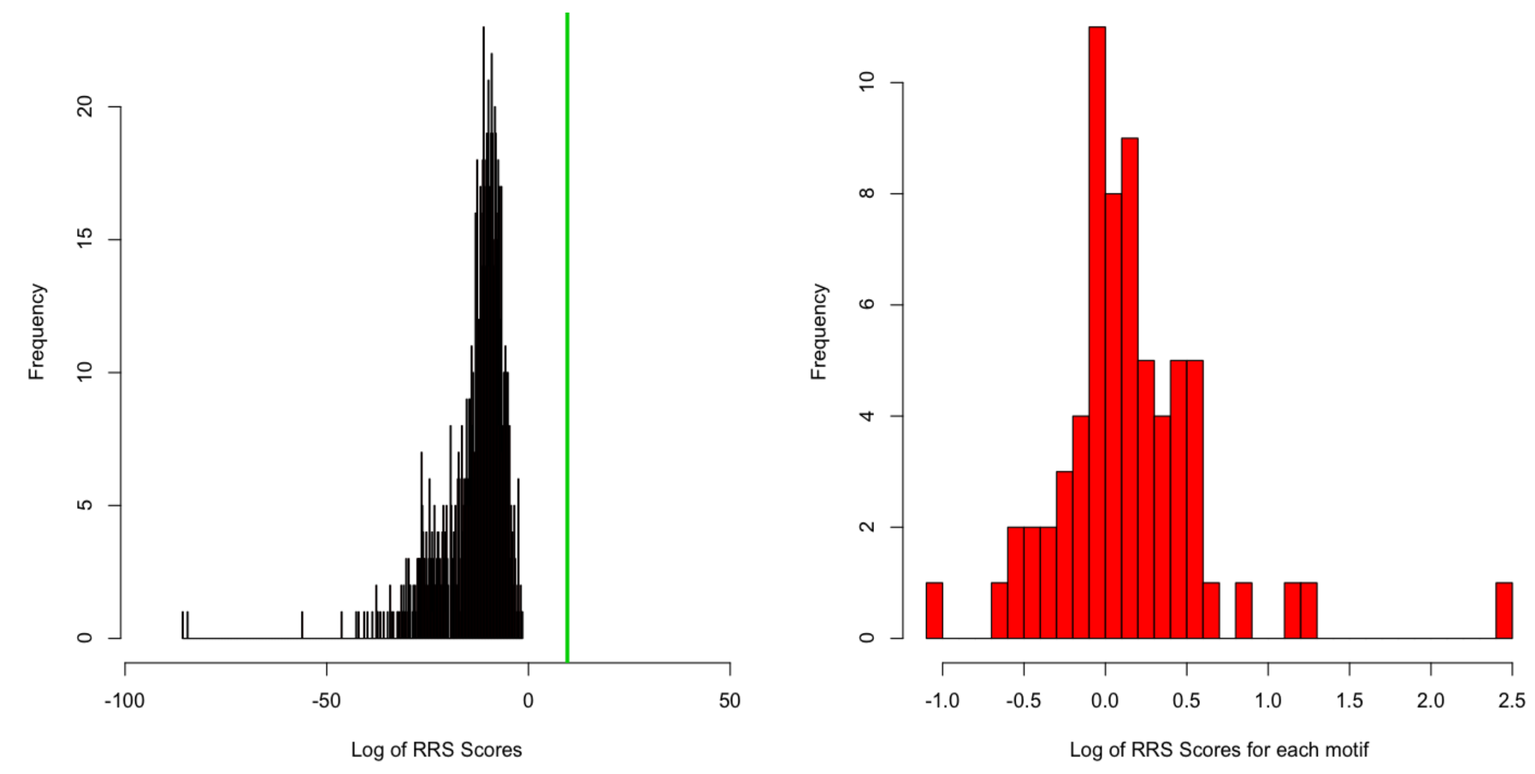
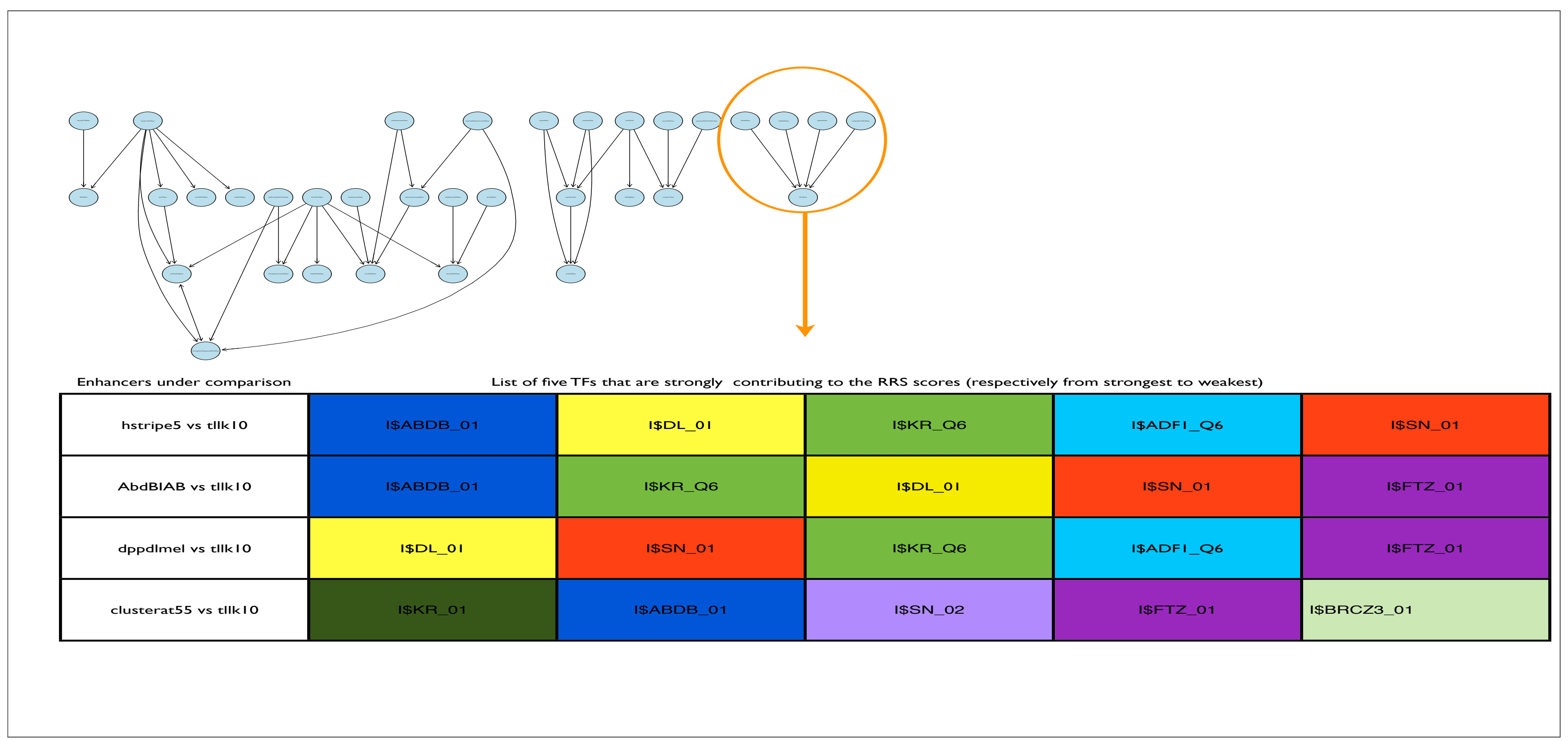


Figure 2: Significance of RRS scores: Left−hand side is illustrating the significance of RRS score of BLASTODERM eve_stripe1 vs cocotd186 (vertical green line) compared to RRS scores of eve_stripe1 vs 1000 randomly picked sequences from the D.melanogaster genome (black histogram). Right−hand side is showing the contribution of 67 insect weight matrices to the score of eve_stripe1 vs cocotd186. According to our model, factors which have scores greater than 1 are strongly contributing (strong presence) to functional similarity of the enhancers. In this example factors BCD, KR and FTZ (in this order) have the highest scores. Interestingly, factor SRY_β is consistently getting a score below −1 in both enhancers, which means strong absence of this factor.



Enhancers under comparison	List of five TFs that are strongly contributing to the RRS scores (respectively from strongest to weakest)				
hstripe5 vs tllk10	I$ABDB_01	I$DL_01	I$KR_Q6	I$ADF1_Q6	I$SN_01
AbdBIAB vs tllk10	I$ABDB_01	I$KR_Q6	I$DL_01	I$SN_01	I$FTZ_01
dppdlmel vs tllk10	I$DL_01	I$SN_01	I$KR_Q6	I$ADF1_Q6	I$FTZ_01
clusterat55 vs tllk10	I$KR_01	I$ABDB_01	I$SN_02	I$FTZ_01	I$BRCZ3_01

Thursday, 18 March 2010

Figure 3: Top graph: Edges are showing 39 top RRS scores from the positive set (131 enhancers). These 39 scores are bigger than any RRS scores from the negative sest (Randomly picked sequences). Nodes are showing the corresponding pairs of enhancers to those 39 scores. Interestingly, this graph is divides to three sub-graphs, suggesting that there might be some common factors regulating each group. The bottom table is giving more details about one of the subgraphs i.e. it is illustrating 5 motifs that are strongly contributing to each of the similarities according to the RRS. Common factors in each pairs of comparisons are color coded.

5 Summary

- RRS is detecting functionally similar sequences from different sets, suggesting that there is a set of common factors regulating those different tissue-specific enhancers.
- Detection of functional conservation of non-alignable sequences.
- With only 67 PWMs as input set of TF motifs which is almost one percent of all 6-mers in the D2Z model, RRS is able to perform comparably to D2Z.
- In comparison to existing similar methods, the RRS is less data-dependent and users may not need any parameter fitting.

References

[1] M. R. Kantorovitz, G. E. Robinson, and S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–55, 2007.

[2] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman. Functional evolution of a cis-regulatory module. *PLoS Biol*, 3(4):e93, 2005.

[3] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–40, 2008.