Alcom-FT Technical Report Series
ALCOMFT-TR-03-99

Motif statistics on plants datasets ⋆

Magali Lescot[1] and Mireille Régnier[2]

[1] LGPD-IBDM, Parc Scientifique de Luminy, CNRS Case 907
13288 Marseille Cedex 9-FRANCE
[2] Mireille.Regnier@inria.fr
Phone: 33 1 39 63 54 78
Fax: 33 1 39 63 55 96
INRIA Rocquencourt, Domaine de Voluceau B.P. 105,
78153 Le Chesnay - FRANCE

Abstract. Mathematical expressions to estimate the probabilities of rare or frequent words in genomic texts. have recently been obtained by different authors. We present a few extensions for the p-value computation, in the Markov probability model: double strand counting, some degenerated consensus. We compare with the results obtained by RSA-tools on some plant datasets. We will exhibit the co-occurrence of the G-box, the I-box and a third motif, very close the Ry-element, also called the Sph Box.

Keywords: genomics, regulatory elements, probability

1 Motivation

The statistical methods are widely used [1, 2, 3, 4] for the search of *regulatory signals* on an entire genome [5], or some part of it, or several genomes [6, 7], or the identification of promoters [8, 9]. They rely on a simple basic assumption: an overrepresented (or underrepresented) word, i.e. a word which occurs significantly more (or less) in real sequences than in random ones, may denote a biological functionality. These words may be for example a *cleavage recognition site* for restriction enzymes or polyadenylation, a *protein binding site* such as a *cis-regulatory element...*

Our first goal is to provide new mathematical results on the so-called p-value. They allow to take into account the dependency between two strands and to compute the p-value for some degenerated consensus patterns. Nevertheless, the proof is beyond the scope of this extended abstract. Our second goal is to prove their relevance on few data in Section 3. Notably, we provide a short comparison with *RSA-tools* [10]. A survey and a comparison of existing softwares can be found in [11].

⋆ This research was partially supported by the French Bioinformatic Program, the French IMPG Program, Liapunov Institute and IST Program of the EU under contract number 99-14186 (ALCOM-FT).

2 Significance criteria. Double strand counting

In statistical softwares, the over- or underrepresentation of a motif is assessed by comparing, subject to a suitable probability model, its observed number of occurrences, N_{H}, with its expected number of occurrences, E_{H}, in a random sequence (or a random set of sequences). Several assumptions on the background probability models can be made [12]. In this paper, we assume that the model is Markovian, with an order ranging from 1 to 3, or Bernoulli. One can also define several criteria. A widely used selection criteria is the *Z-score*: $Z(\mathrm{H}) = \frac{N_{\mathrm{H}} - E_{\mathrm{H}}}{\sqrt{V_{\mathrm{H}}}}$. A more sensitive criterium is the so-called p-value. e.g. the *probability to find that observed number* N_{H} *of occurrences* for the signal under study:

$$p - \mathrm{value(H)} = P(Number\ of\ occurrences\ \geq N_{\mathrm{H}})\ .$$

Approximate computations of the p-value can be found in [13, 14]. An exact computation is provided in [15], that reduces to the solution of a polynomial equation; hence, this is very stable numerically. This formula can be extended for a set of two words, or, a set that has a suitable 2-partition [16]. Again, the p-value computation reduces to the solution of a polynomial equation, that now depends on the *determinant* of a 2×2 matrix. We consider two applications in this paper.

Double strand counting Counting motifs is usually performed on one strand only. However, a binding site may occur on either one of the two strands, and, usually, there is not enough information to allow to chose the correct strand. The statistics of the counting on two strands is hard, due to the fact that the two strands are not independent: indeed, they are complementary. Still, we simply observe that the number of occurrences of a word H on two strands is equal to the number of occurrences of H *and* its complement $\tilde{\mathrm{H}}$ on any of the two strands. This leads to a *weighted counting scheme* described for a set of motifs in [17]. Here, we have a set of size 2, that contains two *complementary* words.

Degenerated consensus We now consider sets of words $\mathcal{H}$ that have a *stable partition* [16] of size 2. The intuition is to create a partition such that each subset in the partition can be viewed as a single word. As a practical consequence, the formulae for a set of two words will apply.

Definition 1. [16] A partition $(\mathcal{H}_1, \mathcal{H}_2)$ of a set $\mathcal{H}$ is *stable* if exists sets $(A_{i,j})_{1 \leq i,j \leq 2}$ such that:

$$\forall (i,j), \forall \mathrm{H}_i \in \mathcal{H}_i, A_{i,j} = \cup_{\mathrm{H}_j \in \mathcal{H}_j} A_{\mathrm{H}_i, \mathrm{H}_j}$$

Intuitively, this means that *any* word in $A_{i,j}$, concatenated to *any* word in $\mathcal{H}_i$ creates an overlapping occurrence with a word in $\mathcal{H}_j$.

I-box The I-box is degenerated, on the 5-th and 6-th positions, and the possible occurrences can be rewritten in IUPAC code as:GATAHB, where H represents A,C or T and B represents C,G or T. The partition of size 2 defined by: $\mathcal{H}_1 =$ GATAHG and $\mathcal{H}_2 =$ GATAHY, where Y stands for C or T is stable. One has:

$$
\begin{aligned}
A_{1,1} &= \{\epsilon, ATAHG\} \\
A_{1,2} &= \{\epsilon, ATAHY\} \\
A_{2,2} &= \{\epsilon\} \\
A_{2,1} &= \emptyset
\end{aligned}
$$

RY-element Let $\mathcal{H}$ be the degenerated word $X.GCATGCA$, where $X = \{AT, AG, TT, GT, TG, TA\}$. The partition $X_1 = \{AT, AG\}, X_2 = \{TT, GT, TG, TA\}$ defines a stable partition of $\mathcal{H}$, with $\mathcal{H}_1 = AKGCATGCA$ and $\mathcal{H}_2 = X_2.GCATGCA$. One has:

$$
\begin{aligned}
A_{1,1} &= \{\epsilon, KGCATGCA\} \\
A_{1,2} &= \emptyset \\
A_{2,2} &= \{\epsilon\} \\
A_{2,1} &= KGCATGCA
\end{aligned}
$$

We realized a Maple implementation of our formulae. We tested them to extract the G-box, I-box or the *prolamin*-box from promoter sequences of plant regulons. This approach is more sensitive than RSA-tools, that uses a very rough approximation to take into account the two strand dependency.

3 Results on plants datasets

Three datasets based on dicot plant sequences were collected using the query system of PlantCARE database (http://intra.psb.ugent.be:8080/PlantCARE/) described in [18] and the sequence retrieval tool of INCLUSive [19] available at (http://www.esat.kuleuven.ac.be/ dna/BioI/Software.html). They are sets of upstream sequences of genes regulated by a same transcription factor (also called regulon). The size of each sequence is up to 800 bp. In pattern discovery the redundancy of sequences in the data set is an important problem. Related sequences could strongly bias the probabilistic calculation by counting twice all the oligonucleotides of the same sequence. The program Purge_sequence from Stefan Kurtz (installed at the web site of RSA-tools http://rsat.ulb.ac.be/rsat/) allowing to remove the repeated regions in the sequences was used first. Once the datasets purged, the program Oligo-analysis [20] from RSA-tools was ran on the different plant regulons. The expected frequency for each word was calculated on the basis of observed subword frequencies, according to a Markov chain model [21]. QuickScore p-value was computed for these top scoring words.

H1	H2	P-value	P-rank	RSA-rank	RSA-Value
cacgtg	cacgtg	0.006025877172	1	1	18.97
atatat	atatat	0.003409070579	2	2	14.51
tatata	tatata	0.002990704076	3	3	14.33
aaaaaa	tttttt	0.001950411598	4	5	11.72
cttatc	gataag	0.001755126172	**5**	24	0.48
acgtgg	ccacgt	0.001303831173	6	4	13.61
gcatgc	gcatgc	0.001133615113	7	28	0.05
agaaaa	ttttct	0.001001527606	8	13	2.32
cgtggc	gccacg	0.0009731707396	9	9	5.50
ctcttc	gaagag	0.0008981282100	10	11	4.02
aagaaa	tttctt	0.000763203980	11	17	0.98
catgca	tgcatg	0.0007546907283	12	10	5.22
agagaa	ttctct	0.0007301625055	13	14	2.11
agagag	ctctct	0.0006848331484	14	6	8.93
aagaga	tctctt	0.0006195342872	15	15	1.70
acacgt	acgtgt	0.0006030184508	16	7	6.53
gtggca	tgccac	0.0005156594314	17	16	1.68
gagaga	tctctc	0.0004850796689	18	8	5.66
aaaata	tatttt	0.000467987509	19	25	0.38
agaggg	ccctct	0.0003809071942	20	18	0.88
cgtgtc	gacacg	0.0003118252875	21	26	0.09
agaaga	tcttct	0.0002925486336	22	22	0.55
cttctc	gagaag	0.0002274952859	23	20	0.74
gagaaa	tttctc	0.0002137461048	24	27	0.08
agagtg	cactct	0.0001810895985	25	23	0.50
aatata	tatatt	0.0001044411667	26	12	2.67
atatta	taatat	0.0000341865540	27	19	0.86
ataata	tattat	0.0000302504021	28	21	0.62

Table 1. G-box: RSA-tools and QuickScore under the Bernoulli model.

Columns 3 and 4 yield $-\log(p - \text{value})$ and the rank from QuickScore
Columns 5 and 6 yield the rank and the score from RSA-tools.

G-box regulon This regulon contains 31 promoter sequences having at least one binding site per sequence recognized by the G-box Binding Protein (GBP). The G-box (CACGTG) is a very well conserved regulatory element [22]. The description of G-box dataset is available in [23] or at (http://www.psb.rug.ac.be/bioinformatics/lescot/Datasets/ListG-boxes.html). It turns out from the results depicted in Table 1 that the Bernoulli model is good enough to extract the G-box, which is a very strong signal. The RSA-heuristics and QuickScore both find it at the top. It is worth noticing that, using QuickScore, the second signal that is not a TATA-like motif is the I-box, GATAAG, that is also in this data set. An other interesting result is the occurrence in the seventh position of the palindromic motif GCATGC that was found only at the 28th position with Oligo-analysis. The significance of these co-occurrences is discussed in Section 4 below.

We studied the degenerated *I*-box, GATAHB, that occurs 59 times in this regulon. Interestingly, its expectation is 67 in this data set, which means that the degenerated *I*-box is *underepresented*. Taking into account this degeneracy, that was observed on other data sets, introduces noise here.

H1	H2	*P*-value	*P*-rank	RSA-rank	RSA-Value
tatata	tatata	0.01845856718	1	1	13.92
atatat	atatat	0.01629054226	2	3	8.68
aaaata	tatttt	0.002138020588	3	11	1.15
agataa	ttatct	0.002102778614	4	10	1.45
cacgtg	cacgtg	0.002053834591	5	9	1.89
cttatc	gataag	0.002014702806	6	5	3.29
ccggcc	ggccgg	0.001634645437	7	7	1.95
cgtggc	gccacg	0.001220101817	8	12	0.80
cacaca	tgtgtg	0.0005046176607	9	4	4.42
acgtgg	ccacgt	0.0004544684399	10	6	2.35
ccacac	gtgtgg	0.0002476596548	11	15	0.11
agagaa	ttctct	0.0002387894225	12	18	0.02
cttctc	gagaag	0.0001492306044	13	14	0.21
ctcctc	gaggag	0.0001151632612	14	16	0.10
gtgtga	tcacac	0.0000798594818	15	17	0.04
aacaag	cttgtt	0.0000379828083	16	8	1.93
aaaaaa	tttttt	0.0000191179417	17	2	9.26
acaaag	ctttgt	0	18	13	0.22

Table 2. I-box: RSA-tools and QuickScore.

.

I-box regulon The second regulon was built based on 17 sequences containing an I-box (GATAAG). This regulatory element, compared to the G-box, is considerably less conserved (more degenerated). Table 2 shows the results for a Markov model of order 1, using QuickScore and RSA-tools. The known regulatory sites are found among the top seven words (against the top ten words using Oligo-analysis). The word AGATAA (overlapping with the *i*-box) went up in the list from the 10th to the 4th place, the *I*-box GATAAG goes down to 6-th rank from its 5-th rank and the G-box (CACGTG) is now in the 5th position (against the 9th). These 3 *p*-vslues are very closed. Among the top seven words, the TATA-box is still in the first position and the second and third words are TATA-like sites. The next signal, AACAAG, has a significantly smaller value (0.0004 yields a 10^{-24} factor). These results are consistent with what is known about the sequences used to make the dataset. For instance, G-box regulatory sites have been described in some of the sequences of the *I*-box dataset. Moreover as the sequences were extracted upstream of the transcription site of the gene, we did expect that these sequence contain a TATA-box.

Prolamin-box regulon A set of 17 sequences was constituted with a collection of *prolamin-box* which is the binding site of the prolamin-box binding factor (PBF). The conserved motif is TGTAAAG.

H1	H2	P-value	P-rank	RSA-rank	RSA-Value
tccaaa	tttgga	0.001520395849	1	3	2.29
ctttac	gtaaag	0.001398337504	2	1	3.69
agatga	tcatct	0.001315937051	3	2	2.57
tgtaaa	tttaca	0.001212588854	4	7	1.32
gactca	tgagtc	0.001179388620	5	5	1.51
gagtca	tgactc	.001179388620	5	6	1.51
ccaaaa	ttttgg	0.001142295768	7	8	1.31
agtcat	atgact	0.001074979528	8	11	0.57
catgac	gtcatg	0.001069876948	9	16	0.03
catggc	gccatg	0.001065492336	10	15	0.17
agatca	tgatct	0.000835385162	11	9	1.05
cttgta	tacaag	0.000728024223	12	10	1.05
aaaagc	gctttt	0.000712743838	13	12	0.50
aaaggt	accttt	0.000646761719	14	14	0.17
atcatc	gatgat	0.000571133009	15	13	0.18
tataaa	tttata	0.000372483347	16	4	2.04

Table 3. prolamin-box: RSA-tools and QuickScore under the Bernoulli model.

Although GTAAAG leaves its 1st rank to the irrelevant TCCAAG (3rd with RSA), the signal TGTAAA goes up to 4th rank from 7th. A further study on heptamers improves this slight drawback and allows to merge these two signals in a larger signal, that actually is the prolamin box. As a matter of fact, 15 out of the 18 occurrences of GTAAAG are preceded by a T. In a double strand counting, TGTAAAG and its complementary word CTTTACA occur 19 times in our data set, while their expectation is 2.012682438. It turns out that the p-value is $.2736992557e-2$, which is higher than all p-values for hexamers, and notably higher than the p-value of GTAAAG.

4 Validation

We try and assess the significance of the new signals that appeared in our data sets.

RY-element We try here and check whether the overrepresented motif GCATGC could be part of the RY-element described in the literature [24, 25, 26], also called a Sph box [27]. Recently the co-occurence of the G-box and RY-elements has been identified in the promoter of the *phas* gene in *Phaseolus vulgaris* [26]. This RY-element is described by a short consensus sequence, CATGCA, or a long sequence, that is either CATGCATG [26] or {G/C}CATGCNN{G/C} [24]. In all these works - and in other references available in PlantCARE database- show that this cis-regulatory element is involved in seed-specific regulation.

In our data set, the motif CATGCATG occurs only twice, in the sequence of gene AJ224116 of *Helianthus annuus*. Indeed, this regulatory element has been described in this promoter sequence (data from PlantCARE).

CCAC-box The third signal is CCACGT. It can be viewed as a CCAC-motif described in [24] as co-occurring with the RY-element in the *Phaseolus vulgaris dlec2* promoter. It can also be viewed as an artefact of the G-box CACGTG. We clarify this point with some word counts and a p-value computation.

The expectation of CCACGTG is 1.6. It occurs 13 times and 10 occurrences are followed by a G, that is to say overlap a *G*-box. Hence, the overrepresentation of CCACGT is explained by the overrepresentation of the G-box CACGTG, which is preferentially preceded by a C, or an A (11 occurrences).

Still, CCAC is also overrepresented in our dataset. Its expectation is 29, it occurs 76 times, and only 10 out of these 76 occurrences overlap a *G*-box. Its QS-value is 0.0013257, which is very closed to the value given for CCACGT, which is 0.0013038.

References

1. A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Prediction of Regulatory Elements in Silico on a Genomic Scale. *Genome Research*, 8:1202–1215, 1998.
2. H. Bussemaker, H. Li, and E. Siggia. Building a dictionary for genomes:Identification of presumptive regulatory sites by statistical analysis. *PNAS*, 97(18):10096–101000, 2000.
3. X. Liu, D.L. Brutlag, and J. Liu. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed gene. In *6-th Pacific Symposium on Biocomputing*, pages 127–138, 2001.
4. E. Eskin and P. Pevzner. Finding Composite Regulatory Patterns in DNA Sequences. *Bioinformatics*, 1(1):1–9, 2002.
5. J. van Helden. Prediction of transcriptional regulation by analysis of the non-coding genome, 2003. in press.
6. M. Gelfand, E. Koonin, and A. Mironov. Prediction of transcription regulatory sites in *archaea* by a comparative genome approach. *Nucleic Acids Research*, 28:695–705, 2000.
7. U. Ohler, G. Liao, H. Niemann, and G. Rubin. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genetics*, 17:56–60, 2001.

8. M. Markstein, P. Markstein, V. Markstein, and M. Levine. Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the drosophila embryo. *PNAS*, 99(2):763–768, 2002.
9. U. Ohler, G. Liao, H. Niemann, and G. Rubin. Computational analysis of core promoters in the drosophila genome. *Genome Biology*, 3(12):0087.1–0087.12, 2002.
10. J. van Helden, B. Andre, and J. Collado-Vides. A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16:177–187, 2000.
11. M. Lescot. *Analyse bioinformatique des régions régulatrices d' Arabidopsis Thaliana et de Drosophila Melanogaster.* Thèse, Université de la Méditerranée, Marseille, 2002. Defended December, 10-th, 2002.
12. S. Hamspon, D. Kibler, and P. Baldi. Distribution patterns of over-represented k-mers in non-coding yeast dna. *Bioinformatics*, 18:513–528, 2002.
13. G. Reinert, S. Schbath, and M. Waterman. Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology*, 7(1):1–46, 2000.
14. G. Bejerano. Efficient exact value computation and applications to biosequence analysis. In M. Vingron, S. Istrail, P.A. Pevzner, and M. Waterman, editors, *RECOMB'03*, pages 38–47. ACM Press, 2003. Proc.RECOMB'02, Berlin-York.
15. M. Régnier and A. Denise. Rare Events on Random Strings, 2003. submitted;http://algo.inria.fr/regnier/index.html.
16. M. Régnier. Complexity of Unusual Words Counting. In *JOBIM'00*, volume 2066 of *Lecture Notes in Computer Science*, pages 101–117. Springer-Verlag, 2001. Proceedings of JOBIM'00, Montpellier.
17. M. Régnier, A. Lifanov, and V. Makeev. Three variations on word counting. In *GCB'00*, pages 75–82. Logos-Verlag, 2000. Proc. German Conference on Bioinformatics, Heidelberg; submitted to BioInformatics.
18. M. Lescot, P. Dehais, G. Thijs, K. Marchal, Y. Moreau, Y. Van de Peer, P. Rouze, and S. Rombauts. Plantcare, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.*, 30:325–327, 2002.
19. G. Thijs, Y. Moreau, F. De Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouze, B. De Moor, K. Marchal, P. Dehais, and Y. Van de Peer. Inclusive: Integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, 18:331–332, 2002.
20. J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.
21. J. van Helden, M.del Olmo, and J. E. Perez-Ortin. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, 28:1000–1010, 2000.
22. R. G. Donald and A. R. Cashmore. Mutation of either g box or i box sequences profoundly affects expression from the arabidopsis rbcs-1a promoter. *Embo J.*, 9:1717–1726, 1990.
23. G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17:1113–1122, 2001.
24. A. Bobb, M.S. Chern, and M. Bustos. Conserved RY-elements mediate transactivation of seed-specific promoters by the developmental regulator PvALF. *Nucleic Acids Research*, 25:641–647, 1997.
25. T. Urao, K. Yamaguchi-Shinozaki, N. Mitsukawa, D. Shibata, and K. Shinozaki.

Molecular cloning and characterization of a gene that encodes a MYC-related protein in Arabidopsis. *Plant Mol Biol.*, 32(3):571–576, 1996.
26. MB. Chandrasekharan, KJ Bishop, and TC Hall. Module-specific regulation of the beta-phaseolin promoter=20 during embryogenesis. *Plant J*, 33(5):853–866, 2003.
27. T.Hattori, Terada, and S. Hamasuna. Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of cis-acting promoter elements required for regulation by abscisic acid and VP1. *Plant J*, 7:913–925, 1995.