Alcom-FT Technical Report Series
ALCOMFT-TR-01-158

Three variations on word counting

Mireille Régnier and Alexandre Lifanov and Vsevolod Makeev
INRIA, 78153 Le Chesnay, FRANCE
Engelhardt Institute of Molecular Biology (EIMB), Vavilov 32, Moscow 117984, RUSSIA *

Abstract. We address the problem of assessing statistical significance of pattern occurrence frequency in biopolymer sequences; to this end we investigate the distribution of pattern occurrences in random sequences, under Bernoulli and Markov model.

We demonstrate how explicit known formulae for expectation and variance are modified in different pattern counting schemes customary in the computational biology, and provide dedicated fast procedures. We discuss several approximations, notably the widely used approximation of the variance by the expectation. Finally, we discuss sensitivity of statistical evaluation to the counting scheme or the probability model. A special attention is paid to changes in Z-scores that may introduce false positive. We provide new criteria to estimate *a priori* validity of various approximations or sensitivity, as well as simple optimized procedures.

We consider three applications. First, several searching algorithms for regulatory sites make use of statistics on consensus words in a single-stranded DNA text. Second, we propose a new counting scheme to count consensus words on both strands in double stranded DNA. Our third application is the counting of profiles, especially PROSITE regular expressions.

1 Introduction

1.1 Biological motivation

Detection of frequent or rare patterns in sequences of biopolymers is a fundamental technique in computational biology (Waterman, 1989). A pattern is a set of words complying with some requirements (PROSITE motif, consensus sets,...). To assess that a given pattern is avoided or overrepresented, one needs a theoretical model for character distribution in sequences and a statistical tool to compare data to the model. We consider below Bernoulli and Markov random models for sequences. Any reasonable statistical tool, notably the Z-score, makes use of the mean and variance. This recently motivated many works. In (Régnier, 2000b), we provide a method for mean and variance computation that is valid for Bernoulli and Markov models and

*This research was supported by IST Program 99-14186 (ALCOM-FT), REMAG Action from INRIA, INTAS grant 99-1476 and MGRI project 244.

several counting constraints (for example, overlap may or may not be allowed). The aim of this paper is twofold. First, we demonstrate that such calculations can be practical regarding their time and space consumption. We provide some modifications of general formulae derived in (Régnier, 2000b) for specific cases relevant in computational biology and practical optimized computation procedures

* We first address statistical significance of a *consensus* word (Berg & Hippel, 1987; van Helden *et al.*, 1998; Ponomarenko, 1999). Computational efficiency is crucial for consensus identification, where statistical significance is checked for a huge number of motifs. This part extends to structured motifs (Marsan & Sagot, 2000).

* Bias appears in overrepresentation criteria when a binding site is searched on *both strands*, due to a lack of biological information. Counting scheme should be weighted.

* Characteristic profiles (Hart *et al.*, 2000; Corpet *et al.*, 2000) for a protein family may be a *regular expression*, for instance in PROSITE format.

All over the paper, we establish a validity domain for some widely used approximations, and discuss sensitivity of statistical tests to Bernoulli approximation to Markov model. This is the second aim of our paper. In passing, we provide simple and efficient choice criteria, based on approximation result. This allows to avoid complex computation of a more intricate formula whenever approximation is tight.

1.2 Mathematical background

General formulae for mean and variance can be found in (Régnier & Szpankowski, 1997; Régnier, 2000b). For sake of simplicity, we state them here in the Bernoulli model, and assume that a pattern is made of words of equal length m.

Notation: Given a pattern H, one denotes $P(\mathrm{H}) = \prod_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{p}_{\mathrm{H_i}}$. Given a set $\mathcal{H}$, one denotes $P(\mathcal{H}) = \sum_{\mathrm{H}\in\mathcal{H}} P(\mathrm{H})$. Denote n the size of the random text.

Fact: Given a set $\mathcal{H}$ of patterns, the expectation of the number of $\mathcal{H}$-occurrences in a random text of size n generated by a Bernoulli model, $N_{\mathcal{H}}$, is:

$$E(N_{\mathcal{H}}) = (n - m + 1)P(\mathcal{H}) \ . \tag{1}$$

It turns out that variance depends on the so called *correlation sets* introduced in (Guibas & Odlyzko, 1981) for the Bernoulli model.

Definition 1 *Given two strings H and F, the* overlapping set *is the set of* H-*suffixes that are* F-*prefixes. Associated* F-*suffixes form the* correlation set $\mathcal{A}_{\mathrm{H,F}}$. *One denotes:*

$$A_{\mathrm{H,F}} = \sum_{w\in\mathcal{A}_{\mathrm{H,F}}} P(w) \ .$$

Intuitively, a word in $\mathcal{A}_{\mathrm{H,F}}$, when concatenated to H, creates an (overlapping) occurrence of F. For example, let H $= 11011$ and $F = 1110$. Then $\mathcal{A}_{\mathrm{H,F}} = \{10, 110\}$ and $\mathcal{A}_{\mathrm{F,H}} = \{11\}$. Associated probabilities are, in biased Bernoulli model where $(p_0, p_1) = (1/3, 2/3)$, $P(A_{11011,1110}) = \frac{2}{9} + \frac{4}{27}$ while $P(A_{1110,11011}) = \frac{4}{9}$.

In the overlapping model, matricial expressions provided in (Régnier, 2000b) for the variance and covariance reduce to:

Theorem 1.1 *(Régnier & Szpankowski, 1997) Given a set $\mathcal{H}$ of patterns, the variance of the number $N_{\mathcal{H}}$ of $\mathcal{H}$-occurrences is, in the Bernoulli model:*

$$Var(N_{\mathcal{H}}) = nc_1 + c_2 \tag{2}$$

where $c_1 = P(\mathcal{H})(1 + (1 - 2m)P(\mathcal{H})) + 2\sum_{\mathrm{H}\in\mathcal{H}} P(\mathrm{H})(\sum_{\mathrm{F}\in\mathcal{H}} \mathrm{A}_{\mathrm{H,F}} - 1)$ and $c_2 = O(m/n)$, and n the size of the random text.

1.3 Statement of the problems

From (1), efficient computation of the mean is equivalent to an efficient computation of $P(\mathcal{H})$. We show on our examples, that one can group computations on $\mathcal{H}$ in order to get time complexity much smaller than theoretical complexity: $O(|\mathcal{H}|)$.

Then we discuss widely used approximation $Var(N_{\mathcal{H}}) = E(N_{\mathcal{H}})$. Theorem 1.1 provides two correcting terms. First, $(1 - 2m)P(\mathcal{H})$ may not be negligible with respect to 1. We simply point out that this correcting term is known, whenever mean is computed; as correction has no extra cost, we suggest a systematic correction.

Second term depends on *autocorrelation*. Intuitively, these correlations are rare events. Nevertheless, it is also known, and proved theoretically, that, no matter how big m is or what is the character distribution, exist sets of words with high correlations. We characterize them *a priori* in the case of consensus words. We provide an efficient algorithm to compute the correcting term, *SmartVar*, to be used in that case (only).

Finally, we discuss sensitivity of statistical test to the distribution. As Bernoulli approximation provides (much) simpler computations, one should use it whenever it is possible. Possible drawback is a *shift of the mean*. We discuss a simple criterium that allows to decide whether this shift may possibly change significance, and introduce false positive. Here, we concentrate on Z-score.

2 Consensus words

Regulatory sites or binding sites may be defined by a consensus. Intuitively, a binding site derives from a single word -the consensus- by a set of allowed operations (substitutions, insertions, deletions,...).Below, we consider the case where substitutions are allowed (the so-called Hamming distance).

Definition 2 *A* k-neighbour *of a word* H *is a word* F *derived from* H *through* k *substitutions. The* k*-neighbourhood is the set of its* k*-neighbours and it is denoted by* $B_k(\mathrm{H})$.

Example: Let H = ACCA. Then, $CCCA, ATCA$ and $ACCT$ are in $B_1(\mathrm{H})$ while $ATCC$ is in $B_2(\mathrm{H})$.

Expectation formulae (1) rewrite for a neighbourhood as:

Theorem 2.1 *Let* H *be a pattern of size* m. *Then, in the Bernoulli model,*

$$\begin{aligned} E(N_{B_1(\mathrm{H})}) &= (n-m+1)P(\mathrm{H})(\sigma_1 - \mathrm{m}) \ , \\ E(N_{B_2(\mathrm{H})}) &= (n-m+1)P(\mathrm{H})\frac{\mathrm{m}(\mathrm{m}-1) - 2\sigma_1(\mathrm{m}-1) + \sigma_1^2 - \sigma_2}{2} \ , \end{aligned}$$

where $\sigma_k = \sum_{i=1}^{m} (\frac{1}{p_{\mathrm{H_i}}})^k, k \geq 1$, *and* n *is the size of the random text.*

We skip the proof in this abstract. Let us point out that one needs only computing $P(\mathrm{H})$ and σ_1 (possibly σ_2 or σ_i), which is very fast and numerically stable. These formulae extend to Markovian case.

Remark: In the uniform case, multiplicative factor reduces to $\frac{m(m-1)}{2}(q-1)^2 P(\mathrm{H})$ that is the size of $B_2(\mathrm{H}) - \mathrm{B}_1(\mathrm{H})$.

Variance computation Our extensive computations of variance through exact formulae -including the constant term- confirm that first correcting term is enough in most cases. Second correction is needed for periodic words $w = F^p$. The smaller is F, the larger is the overrepresentation, For these words, one needs to compute correlation term in (2), e.g. $\sum_{\mathrm{F}\in\mathcal{H}} P(\mathrm{F})(\sum_{\mathrm{G}\in\mathcal{H}} \mathrm{A}_{\mathrm{F,G}} - 1)]$. One naive way is to compute all possible correlation sets and their associated expectations. Time and space complexity are $O(|\mathcal{H}|^2)$. Our first implementation showed this actually represents a strong limitation for practical applications. Efficient grouping of correlation sets reduces quadratic complexity of naive algorithm to a *linear complexity*, for a neighbourhood:

SmartVar Algorithm For any F in the neighbourhood, and any position j in $[1, m]$, one checks whether $(m+1-j)$-suffix is a l-neighbour of the $(m+1-j)$-prefix of H, for some $l \leq k$. This only involves character comparisons. If it is, one may concatenate any $(k-l)$-neighbour of $(j-1)$-suffix of H. Expectation of this set is computed through Theorem 2.1 and memorized in a $[k, m]$-array.

Example Let H = abacaba. Assume 3 errors are allowed. Denote $F = abacaaa$. 3-suffix aaa is a 1-neighbour of 3-prefix aba of H. One may concatenate any 2-neighbour of $caba$. Other possible suffixes are aa, with $l = 1$ and a, with $l = 0$. Hence, $\sum_{\mathrm{G}\in\mathcal{H}}(A_{\mathrm{F,G}} - 1) = E(N_{B_2(caba)}) + E(N_{B_2(acaba)}) + E(N_{B_3(bacaba)})$.

Z-score significance Our algorithmic chain is the following. Compute Z-score with approximate formulae. If H is a periodic word F^p, compute $\alpha = E(B_k(\mathrm{F}))$. If a multiplication by $\frac{1}{\sqrt{(1+\alpha)}}$ changes the significance of Z-score, recompute the variance and Z-score through *SmartVar* algorithm above.

3 Two strand

One usual biogical application of the motif counting problem is recognition of sites, to which some molecules specifically bind. Those molecules can be protein factors or hybridized oligonucleotides. Counting motifs in DNA nucleotide sequence is usually performed along one strand only. In reality, however, DNA consists of two complementary strands and binding sites may occur on either one. Not always there is biological information, which allows one to choose the correct strand. This motivates a new counting scheme with the motives counted in both complementary strands. For example, consider double stranded DNA : $(ACAACACGT, TGTTGTGCA)$. The number of occurrences of AC word is four, three in one strand and one in the other (strands are read in opposite directions). This number is equal to the number of occurrences in any of the strands of the $\{AC, GT\}$ motif, where GT is the complementary word. Observe now that self-complementary word ACGT occurs twice in the double-stranded sequence. Therefore, the problem of counting overlapping motifs in double stranded DNA can be reformulated as a problem of weighted counting on a single strand. The motif set is modified by addition of all complementary words. Any word H that had its complementary in initial set -notably any palindromic word- is assigned a weight $w(\mathrm{H}) = 2$. Other words are assigned weight $w(\mathrm{H}) = 1$.

Theorem 3.1 *Let $\mathcal{H}$ be a set of weighted patterns. The weighted number of occurrences is $\tilde{N}_{\mathcal{H}} = \sum_{\mathrm{H}\in\mathcal{H}} w(\mathrm{H})\mathrm{N_H}$ satisfies $E(\tilde{N}_{\mathcal{H}}) = (n-m+1)\sum_{\mathrm{H}\in\mathcal{H}} w(\mathrm{H})\mathrm{P}(\mathrm{H})$ Formula (2) holds with linearity constant c_1:*

$$\sum_{\mathrm{H}\in\mathcal{H}} w(\mathrm{H})^2\mathrm{P}(\mathrm{H}) + (1-2\mathrm{m})(\sum_{\mathrm{H}\in\mathcal{H}} \mathrm{w}(\mathrm{H})\mathrm{P}(\mathrm{H}))^2 + 2\sum_{\mathrm{H}\in\mathcal{H}} \mathrm{w}(\mathrm{H})\mathrm{P}(\mathrm{H})(\sum_{\mathrm{F}\in\mathcal{H}} \mathrm{w}(\mathrm{F})\mathrm{A_{H,F}} - \mathrm{w}(\mathrm{H}))$$

If H is palindromic, then any word in its k-neighbourhood has its complementary word in this k-neighbourhood; hence all words in k-neighbourhood are assigned weight 2, and $E(\tilde{N}_{\mathcal{H}}) = 2E(N_{\mathcal{H}})$ while $Var(\tilde{N}_{\mathcal{H}}) = 4Var(N_{\mathcal{H}})$. Interestingly, Z-score is not changed, and, more generally, is stable with respect to this counting. Our experiments confirmed this fact. Nevertheless, this weighted counting is significant for algorithms that use total number of occurrences as a filter. Remark in passing that brute approximation $V = E$ is very dangerous... We do not repeat here our previous discussion on approximate formulae for variance, but we observe that the general flavour remains: we have a first correcting term at no extra cost, and algorithm for autocorrelation correcting term steadily extends.

4 Z-score sensitivity

We study Z-score sensitivity when Markovian model is approximated by a Bernoulli model. Let α be the relative error between Markovian probability $P_M(\mathcal{H})$ and Bernoulli probability $P_B(\mathcal{H})$. Error on Z-score rewrites as:

$$\Delta Z = \frac{n(P_M(\mathcal{H}) - P_B(\mathcal{H}))}{\sqrt{nP_B(\mathcal{H})}} = \alpha\sqrt{nP_B(\mathcal{H})} \ . \tag{3}$$

For large sequences, a small error α can still lead to significant variation in Z-score.

Pattern profiles Main application here is the automatic discovery of pattern profiles such as PROSITE regular expressions (Hart *et al.*, 2000). We focus on the computation of the mean in the Markovian case.

MeanMarkov is described for rigid patterns of length m, and a Markovian model of order 1 with transition matrix $\mathbb{P}$ and stationary probability vector q. It takes as an input k specified positions $(i_l)_{1\leq l\leq k}$, and associated sets of allowed amino-acids, S_l. Procedure computes Table $chemin[l,x]$ that contains cumulated expectation of $\mathcal{H}$-suffixes of $\mathcal{H}$-occurrences with x at position l. Internal loop is: $chemin[l,x] = \sum_{y\in S_{l+1}} \mathbb{P}_{x,y}^{i_{l+1}-i_l} \times chemin[l+1,y]$.

Example: let $\mathcal{H} = [RK]-[RK]-x-[ST]$. *MeanMarkov* produces $P_M(\mathcal{H}) = \sum_{x\in\{R,K\}} q(x) \sum_{y\in\{R,K\}} \mathbb{P}_{x,y} \sum_{z\in\{S,T\}} \mathbb{P}_{y,z}^2$. Remark, that in Bernoulli model, an unspecified position contributes by a multiplicative factor 1, and one has $P_B(\mathcal{H}) = (P(R)+P(K))^2 \cdot (P(S)+P(T))$.

We claim a tight bound for approximation $Var(N_{\mathcal{H}}) = E(N_{\mathcal{H}})$ can be proved in that case. Intuitively, there are two reasons for that. First, overlapping expressions define rather long suffixes in this context. Second, amino-acid alphabet is huge; hence, each specified amino-acid in such a suffix has a low probability.

To check sensitivity of this model, we estimated nucleotide and dinucleotide frequencies of amino-acids of *Helicobacter pylori* proteom downloaded from TIGR database. We checked ratio of Bernoulli and Markov mean for some PROSITE families. For PROSITE motif $[RK]-[RK]-x-[ST]$, this ratio turns out to be $.1370028784e^{-2}/.1747198482e^{-2}$. Interestingly enough, this would change positive Z-score 22 in (Nicodème *et al.*, 1999) into negative Z- score -6.020723691...(if dinucleotide frequencies are the same in both experiments).

Regulatory sites We checked sensitivity to Bernoulli approximation on upstream regions of genes from *Helicobacter pylori.* For motif TAAAAG, we got $\Delta Z = 4.7$. This definitely is a significative change in Z-scores of order 4.

Rule When expectation is large, compare $\sqrt{E}$ with Bernoulli Z-score, and compute Markov mean when the ratio is "high". To precise this notion, one must specify algorithms, as profile algorithms often chose the "highest score(s)"...

5 Conclusion

Pattern counting in random sequences with known statistical properties has been extensively studied recently, but still new results are being produced. Computational biology, in which discovery of specific patterns is one of the primary problems, provides a natural application for the results of mathematical studies. Recent theoretical results, e.g. exact formulae for mean and variance of a set of words (Régnier, 2000b) allowed us to investigate the relevance of approximations, which are used widely to evaluate statistical significance of patterns found in biopolymer sequences, with the help of Z-score. Although general results are derived through simultaneous use of generating functions and asymptotic analysis, we emphasize that expectation and/or variance can be steadily computed. We provide optimized dedicated procedures, for Bernoulli and Markov models that are an efficient alternative to automaton approach used in (Nicodème *et al.*, 1999).

It turns out that validity of variance approximation by expectation is different in different counting schemes. Notably, we pointed out on double strand counting that it is totally wrong in a weighted counting scheme. Nevertheless, first correcting term provided by exact results is obtained at no extra cost, and we suggest a systematic correction that is precise enough in most cases. However, for any sequence composition, always there are motives, for which second correction is needed. Tightness of approximation depends on words structures: selfoverlap for consensus words (and regular expressions), palindromic structure for both strand counting, and on data bias. We provided simple criteria, to avoid computing more intricate formulae whenever possible.

We also studied sensitivity of Z-score to Bernoulli approximation of Markov model, and pointed out that it may induce a shift of the mean, hence a change of significance. Our experiments were performed on *H. pylori* gene upstream region and proteom. Here again, we provided simple filtering criteria to chose appropriate model, as simple as possible.

Our algorithmic contribution is as follows. In general, complexity of the actual computation of variance depends on autocorrelation automaton (Régnier, 2000a). Simplifications occur for overlapping model addressed here. Our first algorithm, *SmartVar*, provides fast computation of the variance with low space complexity. It applies in more complex contexts, such as expectation computation in the renewal model, or evaluation of the waiting time for a (consensus) word. We also describe a fast algorithm, *MeanMarkov*, to compute expectation of regular expressions. Efficiency of expectation calculation is critical whenever a pattern discovery algorithm (Hart *et al.*, 2000; Corpet *et al.*, 2000) filters out non-significant patterns from a huge set. These functions -to be called by pattern discovery algorithms- are available in C-library *QuickScore* at *http://www.loria.fr/projets/REMAG/logiciels.html.*

References

Berg, O. & Hippel, V. (1987). Selection of DNA binding sites by regulatory proteins. statistical mechanical theory and application to operators and promotors. *J. of Mol. Biol.* **193** (4), 723–750.

Corpet, F., Servant, F., Gouzy, J., & Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267–269.

Guibas, L. & Odlyzko, A. (1981). String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory,* Series A*,* **30**, 183–208.

Hart, R., Royyuru, A., Stolovitzky, G., & Califano, A. (2000). Systematic and Automated Discovery of Patterns in PROSITE Families. In: *RECOMB'00* pp. 147–154, ACM-. Proceedings RECOMB'00, Tokyo.

Marsan, L. & Sagot, M. (2000). Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification. In: *RECOMB'00* pp. 210–219, ACM-. Proceedings RECOMB'00, Tokyo.

Nicodème, P., Salvy, B., & Flajolet, P. (1999). Motif statistics. In: *ESA'99* volume 1643 of *Lecture Notes in Computer Science* pp. 194–211, Springer-Verlag. Proc. European Symposium on Algorithms-ESA'99, Prague.

Ponomarenko, J. (1999). Oligonucleotide frequency matrices addressed to recognizing functional dna sites. *Bioinformatics,* **15**, 631–643.

Régnier, M. (2000a). Complexity of Unusual Words Counting. presented at JOBIM'00;submitted.

Régnier, M. (2000b). A Unified Approach to Word Occurrences Probabilities. *Discrete Applied Mathematics,* . To appear in a special issue on Computational Biology; preliminary version at RECOMB'98.

Régnier, M. & Szpankowski, W. (1997). On the Approximate Pattern Occurrences in a Text. In: *Compression and Complexity of SEQUENCES 1997*, (Society, I. C., ed) pp. 253–264,. In Proceedings SEQUENCE'97,Positano, Italy.

van Helden, J., Andre, B., & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842.

Waterman, M. (1989). *Mathematical Methods for DNA Sequenc.* Boca Raton, Florida: CRC Press Boca Raton, Florida.