BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 18 2007, pages 2488–2490
doi:10.1093/bioinformatics/btm366

Sequence analysis

Quality estimation of multiple sequence alignments by Bayesian hypothesis testing

Andrija Tomovic and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerestrasse 66, CH-4056 Basel

Received on May 17, 2007; revised on July 2, 2007; accepted on July 9, 2007

Advance Access publication July 27, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: In this work we present a web-based tool for estimating multiple alignment quality using Bayesian hypothesis testing. The proposed method is very simple, easily implemented and not time consuming with a linear complexity. We evaluated method against a series of different alignments (a set of random and biologically derived alignments) and compared the results with tools based on classical statistical methods (such as sFFT and csFFT). Taking correlation coefficient as an objective criterion of the true quality, we found that Bayesian hypothesis testing performed better on average than the classical methods we tested. This approach may be used independently or as a component of any tool in computational biology which is based on the statistical estimation of alignment quality.

Availability: http://www.fmi.ch/groups/functional.genomics/tool.htm

Contact: edward.oakeley@fmi.ch

Supplementary information: Supplementary data are available from http://www.fmi.ch/groups/functional.genomics/tool-Supp.htm

1 INTRODUCTION

Statistical estimation of the significance of proposed alignments is one of the central challenges of evaluating the output of all alignment tools. Local ungapped alignments play an important role in the discovery and classification of both DNA and protein sequences. To evaluate a proposed sequence alignment we must know the likelihood of it occurring by chance rather than, for example, deriving from a common ancestral sequence. Statistically significant alignments have a higher chance of being biologically relevant. The evaluation of ungapped local alignment is usually made using its information content or relative entropy (Hertz and Stormo, 1999; Nagarajan *et al.*, 2005):

$$I_{\text{seq}} = \sum_{i=1}^{L} \sum_{j=1}^{|A|} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{b_j} \tag{1}$$

where L is the length of the sequence from an alphabet A, n_{ij} count of the j-th letter in the i-th column of alignment, n is the number of sequences in the alignment and b_j the background frequency of the j-th letter. Using this scoring function (1) and a null model, which assumes that each of the k columns has n letters independently sampled according to the background distribution we can estimate a P-value. The P-value for a given scoring value s_0 represents the probability of an entropy score of s_0 or better under the null model (Hertz and Stormo, 1999; Nagarajan *et al.*, 2005). When the information content (I_{seq}) is small and the number of sequences (n) is large, the value $2nI_{\text{seq}}$ tends to be χ^2-distributed with $k(|A|\text{-}1)$ degrees of freedom (Wilks, 1938). But this approximation is very poor when we have large scores and few sequences, which is a common situation. Several methods have been developed to improve this P-value estimation (Dembo *et al.*, 1994; Hertz and Stormo, 1999; Karlin and Altschul, 1990; Keich, 2005; Nagarajan *et al.*, 2005). In this work, we present a web-based tool for estimating sequence alignment significances without gaps using Bayesian hypothesis testing. Bayesian methods have already been used in algorithms for sequence alignment (Liu and Lawrence, 1999; Liu *et al.*, 1995; Lunter *et al.*, 2005; Suchard and Redelings, 2006; Webb *et al.*, 2002; Zhu *et al.*, 1998), but in our implementation we used a Bayesian approach to evaluate multiple sequence alignments without gaps that had already been generated. This approach can be used independently or as a component of any tool in computational biology which uses statistical alignment quality estimates.

2 METHOD

Quality estimation of multiple sequence alignments by Bayesian hypothesis testing is based on the work (Minka, 1998; Liu and Lawrence, 1999) which we have adapted for use with DNA and protein sequence alignments. In the interest of simplicity, we will demonstrate the utility of this method in the context of DNA sequence alignments, but it can easily be applied to protein sequence alignments too.

Let us define an alignment X of n DNA sequences of length L:

$$\begin{matrix} X_1^1\, X_2^2 \;\ldots\; X_L^1 \\ X_1^2\, X_2^2 \;\ldots\; X_L^2 \\ \ldots \\ X_1^n\, X_2^n \;\ldots\; X_L^n \end{matrix} \tag{2}$$

Let X_i represent the vector frequencies for each letter (base) for column i of a multiple alignment: $X_i = [X(a,i),\ X(c,i),\ X(g,i),\ X(t,i)]$. We also define Y as a vector with the same length as X_i, $Y = [Y(a),\ Y(c),\ Y(g),$

*To whom correspondence should be addressed.

Table 1. Summary of results from the estimation of 207 alignments (100 random and 107 JASPAR-derived) produced by three methods sFFT, csFFT and Bayes method

Method	True positive	True negative	False positive	False negative	Specificity	Sensitivity	Corr. coef.
sFFT	107	60	40	0	0.60	1	0.66
csFFT	107	60	40	0	0.60	1	0.66
Bayes method	97	100	0	10	1	0.91	0.91

$Y(t)] = [y_a n,\ y_c n,\ y_g n,\ y_t n]$, where y_a, y_c, y_g and y_t represents the background frequencies of each base, respectively a, c, g, and t. Background frequencies of each base can be estimated based on input data or user can specify it. To evaluate the alignment (2), first we will test the following hypotheses:

$$\begin{aligned}&\mathrm{H}_0\text{: } Y \text{ and } X_i \text{ come from the same multinomial distribution}\\&\mathrm{H}_1\text{: } Y \text{ and } X_i \text{ come from different multinomial distributions}\end{aligned} \tag{3}$$

This hypothesis testing can be evaluated directly [in a way similar to that described by (Liu and Lawrence, 1999; Minka, 1998), or in the form of a test for independence (Minka, 1998) which gives slightly different results because of different priors. We have used second approach and a detailed description as to how it is possible to convert the hypothesis test (3) into an independence test is given in Supplementary Material 1. For each column we calculated a Bayes factor $BF_i(\mathrm{H_o};\ \mathrm{H}_1)$ and likelihoods $P_i(Y,X_i|\mathrm{H}_0)$ and $P_i(Y,X_i|\mathrm{H}_1)$. Because of our assumption of independence between the columns, after calculating $BF_i(\mathrm{H_o};\ \mathrm{H}_1)$ and $P_i(Y,X_i|\mathrm{H}_0)$ and $P_i(Y,X_i|\mathrm{H}_1)$ for each $i = 1,\dots,L$ (for each column) we can calculate:

$$\mathrm{BF} = \prod_{i=1}^{L} \mathrm{BF}_i(\mathrm{H}_0,\mathrm{H}_1) \tag{4}$$

$$P(\mathrm{H}_0|Y,X) = \frac{P(\mathrm{H}_0)\prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H}_0)}{P(\mathrm{H}_0)\prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H}_0) + P(\mathrm{H}_1)\prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H}_1)} \tag{5}$$

These scores provide us with an estimate of the multiple sequence alignment significance. It is more significant when BF is small (much smaller than 1) and when the posterior probability of the null model $P(\mathrm{H_o}|\ Y,\ X)$ is small (smaller probability of null model for the given alignment, i.e. smaller probability that given alignment is random). Jeffreys' scale (Jeffreys, 1961) of evidence for Bayes factors is given in Supplementary Material 1- Table 2. We used the posterior probability of the random model (null hypothesis) as a final score of alignment quality for the evaluation of our method (see the next section), because it is a more precise score value than Bayes factor.

3 RESULTS AND DISCUSSION

In this section we report our evaluation of the presented method and its comparison to other methods from classical (orthodox) statistics. We took 107 alignments of transcription factor binding sites, representing each factor in the JASPAR database (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004) and calculated the BF (4) and posterior probability of the null hypothesis (random model) (5). Detailed list for each transcription factor and its corresponding posterior probability and Bayes factor is given in Supplementary Material 2. All alignments, but 10, were found to be significant with very small posterior probabilities for the null hypothesis (much smaller than 0.001). Next, we generated 100 random alignments (available from http://www.fmi.ch/groups/functional.genomics/RandomAlignments.zip) using the RSA tool (van Helden, 2003). The random and JASPAR alignments had approximately the same distribution in terms of length and the number of sequences (Supplementary Material 3-Table 1). For each random alignment, we calculated BF (4) and posterior probabilities of the null hypothesis (5) (Supplementary Material 4). All alignments had posterior probabilities higher than 0.99 and they are correctly identified as not being statistically significant (true negatives). There are several classical (orthodox) techniques for the statistical evaluation of local ungapped alignments. Fast, but inaccurate, techniques are used in motif discovery tools [e.g. MEME (Bailey and Elkan, 1994), Consensus (Hertz *et al.*, 1990; Hertz and Stormo, 1999)]. In Supplementary Material 5 - Table 1, we report some of the more accurate methods for the statistical estimation of short ungapped alignments and their running times. The time complexity for the calculation of Bayes factor (4) and posterior probability (5) is linear O(L), and this has advantages over these other methods. We compared results (posterior probabilities of the random model) obtained by Bayesian approach with the *P*-values calculated by two classical methods csFFT (Nagarajan *et al.*, 2005) and sFFT (Keich, 2005) for a the transcription factor binding site alignments of each factor in the JASPAR database and 100 random alignments. In Table 1 we summarize the results for 207 alignments based on the *P*-values provided by the sFFT and csFFT methods, together with the results from the Bayesian method. The calculation of specificity and sensitivity was performed using the following formula:

$$\text{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \quad \text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

Finally, Pearson product-moment correlation coefficients [also called the 'phi coefficient of correlation' (Burset and Guigo, 1996; Tompa *et al.*, 2005)] were calculated using:

$$\text{Corr.Coef.} = \frac{\mathrm{TP^*TN} - \mathrm{FN^*FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})^*(\mathrm{TN}+\mathrm{FP})^*(\mathrm{TP}+\mathrm{FP})^*(\mathrm{TN}+\mathrm{FN})}} \tag{7}$$

Correlation coefficients may take any value between -1 (indicating perfect anticorrleation) and 1 (indicating perfect correlation).

We conclude, based on Table 1, that the Bayesian approach is superior to the classical approaches.

4 CONCLUSIONS

The method for using Bayesian hypothesis tests to evaluate alignment quality is simple, easy to implement and has a linear time complexity. Our method shows very high sensitivity and specificity in distinguishing biologically relevant from random alignments. It performs much better than methods based on classical statistics (Table 1). It can be integrated into any tool that uses statistical estimates of sequence alignments or as a post-processing filter of the output from any tool that returns a number of ordered alignments. Possible applications include: motif finder algorithms; algorithms for profile–profile and sequence–profile alignment; and the analysis of protein domains and their families. Our tool is available at http://www.fmi.ch/groups/functional.genomics/tool.htm.

ACKNOWLEDGEMENTS

We would like to thank Michael Stadler for helpful discussions. This work was supported by the Novartis Research Foundation. Funding to pay the Open Access publication charges was provided by the Novartis Research Foundation FMI.

Conflict of interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Dembo,A. *et al.* (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, **22**, 2022–2039.

Hertz,G.Z. *et al.* (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Jeffreys,H. (1961) *Theory of Probability*. Clarendon Press, Oxford.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Keich,U. (2005) sFFT: a faster accurate computation of the p-value of the entropy score. *J. Comput. Biol.*, **12**, 416–430.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.

Liu,J.S. *et al.* (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Association*, **90**, 1156–1170.

Lunter,G. *et al.* (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83–93.

Minka,T. (2003) Bayesian inference, entropy, and the multinomial distribution. *Technical Report*.

Nagarajan,N. *et al.* (2005) Computing the *P*-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21** (Suppl. 1), i311–i318.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Suchard,M.A. and Redelings,B.D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.

Webb,B.J. *et al.* (2002) BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res.*, **30**, 1268–1277.

Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, **9**, 60–62.

Zhu,J. *et al.* (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.