BMC Bioinformatics

Research article

Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites

Nak-Kyeong Kim, Kannan Tharakaraman, Leonardo Mariño-Ramírez and John L Spouge*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Email: Nak-Kyeong Kim - kimnak@ncbi.nlm.nih.gov; Kannan Tharakaraman - tharakar@ncbi.nlm.nih.gov; Leonardo Mariño-Ramírez - marino@ncbi.nlm.nih.gov; John L Spouge* - spouge@ncbi.nlm.nih.gov

* Corresponding author

Published: 4 June 2008

BMC Bioinformatics 2008, **9**:262 doi:10.1186/1471-2105-9-262

Received: 29 October 2007
Accepted: 4 June 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/262

Abstract

Background: Biologically active sequence motifs often have positional preferences with respect to a genomic landmark. For example, many known transcription factor binding sites (TFBSs) occur within an interval [-300, 0] bases upstream of a transcription start site (TSS). Although some programs for identifying sequence motifs exploit positional information, most of them model it only implicitly and with *ad hoc* methods, making them unsuitable for general motif searches.

Results: A-GLAM, a user-friendly computer program for identifying sequence motifs, now incorporates a Bayesian model systematically combining sequence and positional information. A-GLAM's predictions with and without positional information were compared on two human TFBS datasets, each containing sequences corresponding to the interval [-2000, 0] bases upstream of a known TSS. A rigorous statistical analysis showed that positional information significantly improved the prediction of sequence motifs, and an extensive cross-validation study showed that A-GLAM's model was robust against mild misspecification of its parameters. As expected, when sequences in the datasets were successively truncated to the intervals [-1000, 0], [-500, 0] and [-250, 0], positional information aided motif prediction less and less, but never hurt it significantly.

Conclusion: Although sequence truncation is a viable strategy when searching for biologically active motifs with a positional preference, a probabilistic model (used reasonably) generally provides a superior and more robust strategy, particularly when the sequence motifs' positional preferences are not well characterized.

Background

Transcription factor binding sites (TFBSs) provide a specific example of biologically functional sequence motifs that sometimes have positional preferences. TFBSs contribute substantially to the control of gene expression, and because of their biological importance, much experimental effort has been expended in identifying them. Because experimental identification is expensive, there are now many computational tools that identify TFBSs as the subsequences, or "motifs", common to a set of sequences.

Most TFBSs correspond to short and imprecise motifs [1], however, so all computational tools in a recent contest performed rather poorly in identifying known TFBSs [2].

Although some tools have an *ad hoc* basis [3-5], other tools have a basis in the calculus of probability, and can therefore immediately and systematically combine sequence with other sources of information. Most probabilistic tools align candidate subsequences and convert the nucleotide counts in the alignment columns into a position-specific score matrix (PSSM). Most PSSMs are based on the log ratio between a motif model and a background model. Tools then identify putative motifs by maximizing the log ratio, usually with expectation maximization (EM) [6] or Gibbs sampling [7-9].

Experiments have shown, however, that besides common sequence motifs, TFBSs also have positional preferences, as illustrated in Figure 1. In yeast, TFBS positions demonstrate a strong bias toward locations between 150 and 50 bases upstream of the TSS [10]. In E. coli, TFBS positions tend to be located between 400 and 0 bases upstream of the translation start site [11]. In the words of Wray *et al.*, "for at least some regulatory elements, function constrains their position with respect to the transcriptional start site" (TSS) [1]. On the other hand, the trends regarding the positional preferences of TFBSs appear inconsistent. Wray *et al.* continue "for most transcription factors, however, binding sites lack any obvious spatial restriction relative to other feature of the locus" [1].

Some computational methods do exist to exploit the positional preferences of TFBSs. The first computational study using positional preferences used an empirical prior distribution of known positional information with respect to the translation start site from the *E. coli* genome [12]. This simple method, however, is applicable only to very simple

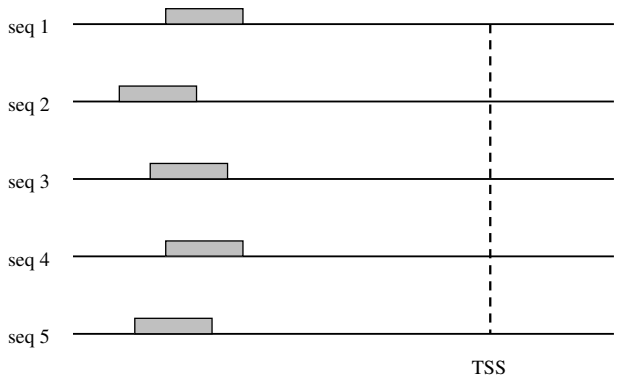

Figure 1
Positions of hypothetical TFBSs (gray boxes) with respect to the corresponding TSS.

organism like *E. coli*. Another computational study used position to calculate p-values for candidate motifs that formed a cluster [13]. The p-values were based on one particular database, however, and might not generalize reliably. Moreover, the corresponding model is not a probability model, making the systematic combination of sequence and positional information problematic. Yet another computational study modeled the positional preferences of TFBSs with a uniform prior, only mentioning the possibility of a more informative prior [11]. A systematic computational study to find new TFBS motifs by exploiting positional preferences applied a chi-square test to bins of positions near TSSs [14]. The chi-square test found one 8-letter word with significant positional preferences, the "Clus1" word, TCTCGCGA. The study's use of binning probably reduced the power of statistical tests, however. Shortly thereafter, in confirmation of the reduced statistical power, a systematic study of a human promoter dataset [15] identified 801 8-letter words with a positional preferences with respect to the TSS [9]. Interestingly, although 388 of the 801 words appeared in the TRANSFAC database [16], 413 of the words did not, suggesting that TFBS positional preferences were much more pervasive than previously believed. A later study showed that in eukaryotes the distribution of TFBSs was not uniform with respect to the TSS [17]. A study using chromatin immunoprecipitation followed by DNA hybridization (ChIP-Chip experiments) inferred TFBSs within sheared DNA fragments by using prior probability distributions to model positional preference [18]. The model was not directed at identifying TFBSs by their positional preferences with respect to genomic landmarks, however. Finally, a study applied a Poisson approximation to bins of positions within promoters to identify TFBSs by their positional preferences with respect to the TSS [19].

Several studies, therefore, have examined the positional conservation of TFBSs. Consequently, TFBS positional preferences are relatively well understood, particularly when compared to most non-coding DNA. Very few computational tools systematically combine positional preference with sequence information, however, and to our knowledge, no general-purpose computational tools using positional information are currently available. Standard tools like MEME [6], AlignACE [10], and MotifSampler [20], e.g., do not use positional information. Accordingly, this article evaluates the accuracy of predictions from a Bayesian model combining sequence with positional information, implemented in the newest version of the tool A-GLAM [9]. We assessed predictions from A-GLAM with and without the positional information, using a standard dataset of sequences with known TFBSs, and were therefore able to measure the contribution of positional information to TFBS prediction accuracy.

Results

Results for the TSS Tompa dataset

The TSS Tompa dataset is one of two test datasets considered in this study and contains 23 data subsets (see Methods). Table 1 shows an anecdotal A-GLAM alignment using positional information for the dataset 'hm08r' from the TSS Tompa dataset, which contains 10 sequences of length 2001. Run in its ZOOPS mode (Zero Or One Per Sequence), A-GLAM returned candidate alignments with only one or zero candidate site per sequence. In addition to sequence conservation, the alignment shows positional conservation within an interval of [-220, -1], much narrower than the input interval, [-2000, 0] bp upstream of the transcription start site (TSS). The alignment also overlapped several known sites (underlined in Table 1), with a correlation coefficient of 0.574, indicating good overlap.

Table 1 does not show the corresponding alignment without positional information, because its width was a biologically unrealistic 126 bp long. The alignment showed little positional conservation, with a range of [-2000, -1237]. It also showed essentially no overlap with the known sites, with a correlation coefficient of -0.012.

For TFBSs predicted without positional information, E-values were immoderately small, even for incorrect predictions. (Some incorrect predictions even displayed a numerical underflow E-value of 0, data not shown.) In contrast, the E-values in Table 1 were quite moderate, perhaps because they had to reconcile conflicting constraints from different sources of information on the motifs.

Alignments for more data subsets can be found in Supplementary Tables 1–6 [see Additional file 1]. We collected alignments (with positional information) whose correlation coefficient (CC) is larger than 0.08. The hm03r data subset does not appear in Tables 1–6, despite a CC of 0.386, because the corresponding alignment had a biologically unrealistic width of 224 bp. Unrealistically large alignment widths are much less common for alignments with positional information than without. In Supplementary Tables 2–6, the alignments without positional information are omitted because they show essentially no overlap with known binding sites.

Table 1: The A-GLAM output with positional information for 'hm08r'.

Name	Start	Alignment	End	Score	E-value
seq_0	-66	GTCACGGC	-59	11.0093	6.65E-06
seq_2	-65	GTGACGTT	-58	10.3315	2.30E-05
seq_3	-58	ATGACGTC	-51	11.2688	2.94E-06
seq_5	-188	GTGACGTC	-181	11.4594	1.28E-06
seq_7	-184	CTGACGAC	-177	9.86871	4.64E-05
seq_9	-101	ATGACGTC	-94	10.9283	8.09E-06
seq_10	-220	ATCACGGC	-213	7.58906	3.78E-04
seq_11	-80	GTGACGTC	-73	11.1306	4.75E-06
seq_12	-52	CTGACGGC	-45	10.0764	3.50E-05
seq_14	-8	CTGATGTC	-1	7.60515	3.69E-04

A-GLAM predicted TFBSs in 10 data subsets in the TSS Tompa data subset hm08r'. The column "Name" shows each data subset; the column "Alignment", the corresponding predicted TFBS. The start and end positions with respect to the corresponding TSS are shown in the columns "Start" and "End". The columns "Score" and "E-value" show bit scores and E-values that A-GLAM assigned to predicted TFBSs. The known binding sites in the alignment are underlined.

Table 2 summarizes results for all 23 TSS Tompa data subsets. Some 18 out of the 23 datasets show improved predictions after adding positional information. Overall, the combined correlation coefficient (CCC; see Methods) at the bottom of Table 2 improved from -0.008 to 0.101. To evaluate the statistical significance, let γ_- and γ_+ denote the average correlation coefficient for each data subset without and with positional information. A one-sample Wilcoxon test against the one-sided null hypothesis $\gamma_- \geq \gamma_+$ yielded a p-value of 0.002, supporting the alternative hypothesis that $\gamma_- < \gamma_+$.

Results for TRANSFAC dataset

The TRANSFAC dataset contains 82 data subsets. Supplementary Table 8 contains detailed results for the input interval of [-2000, 0]. With the addition of positional information, the CCC has improved from -0.009 to 0.027 with a p-value of 10^{-8} (Wilcoxon test as above). The CCC for TRANSFAC dataset (0.027) is smaller than for TSS Tompa dataset (0.101), and the positional information makes a more significant change in the CCC for the TRANSFAC dataset ($p = 10^{-8}$) than for the TSS Tompa dataset ($p = 0.002$), probably because the TRANSFAC dataset contains 82 data subsets; the TSS Tompa dataset, only 23. In the case of subtle differences, the larger TRANSFAC dataset provides more evidence, leading to smaller p-values.

Cross-validation using TSS Tompa dataset

Because we used known binding sites to estimate the hyperparameters of the model (see Methods), one might suspect over-fitting. Moreover, because the distribution of locations might vary from one type of TFBS to another, the proposed model might not be appropriate for the discovery of unknown binding sites of different types of TFBSs. Cross-validation addressed these issues (see Methods).

Over the 100 random partitions from TSS Tompa dataset, the sample average of the CCC was 0.086; its sample standard deviation, 0.027; its 90% confidence interval, (0.049, 0.133); and its range, (0.029, 0.155). (The TRANSFAC dataset was not used for 5-fold cross-validation because of amount of computation required.) The

Table 2: The correlation coefficients for the TSS Tompa data subsets

Data Subset	Without positional information	With positional information	Improvement
hm01r	-0.012	-0.007	0.005
hm02r	-0.009	-0.007	0.002
hm03r	-0.037	0.386	0.423
hm04r	-0.008	-0.005	0.003
hm05r	-0.031	-0.019	0.012
hm06r	-0.014	0.156	0.170
hm07r	-0.015	-0.015	-0.001
hm08r	-0.012	0.574	0.586
hm09r	-0.011	0.358	0.369
hm10r	-0.019	0.083	0.102
hm11r	-0.028	-0.012	0.016
hm13r	-0.015	-0.016	-0.001
hm14r	0.204	-0.018	-0.222
hm15r	-0.011	-0.012	-0.002
hm16r	-0.011	-0.006	0.005
hm17r	-0.015	-0.012	0.004
hm18r	-0.018	0.094	0.112
hm19r	-0.010	-0.007	0.003
hm20r	-0.026	0.046	0.073
hm21r	0.401	0.384	-0.016
hm22r	-0.020	-0.020	0.000
hm24r	-0.016	-0.010	0.006
hm26r	-0.016	0.099	0.115
Combined CC	-0.008	0.101	0.109

Table 2 shows the correlation coefficients for A-GLAM's predictions on the 23 subsets of the TSS Tompa dataset. The column, "Improvement", quantifies the effect of positional information on predictions, by showing the difference between the correlation coefficients in the second and third columns, "Without Positional Information" and "With Positional Information".

CCC for the model using sequence information alone was -0.008. Because the CCC for sequence alone lay outside the range (0.029, 0.155) of the 100 CCCs using positional information in the 5-fold cross-validation, positional information improved prediction accuracy significantly. The actual CCC for the model using both sequence and positional information was 0.101 (see Table 2), well within the 90% confidence interval from cross-validation. The different types of known sites have quite diverse distributions (see Fig. 2), so we expect occasional misspecification of hyperparameters η in our model (see Methods). The 5-fold cross-validation shows, however, that classification accuracy is not excessively sensitive to the hyperparameter estimation or, by extension, to the locations of the known sites.

Truncation effect on sequences of test datasets

Figures 2 and 3 suggest that a truncated input sequence interval of, say, [-500, 0] or [-250, 0] might incorporate positional information as well as a Bayesian positional model applied to the full interval [-2000, 0]. Accordingly, in addition to the full interval [-2000, 0], we tested 3 truncated intervals [-1000, 0], [-500, 0], and [-250, 0]. (See Supplementary Table 7 and 8 for details.) The predictive accuracy, as represented by the CCCs in Table 3, indicate that truncation on its own, without any Bayesian positional modeling, improved the motif predictions. Moreover, predictive improvements due to modeling position gradually disappeared as the truncation reduced the interval to [-250, 0]. Note, however, that positional modeling never significantly hurt the predictive accuracy, even with truncated input sequences.

Discussion

The new version of the A-GLAM program ('anchored gapless local alignment of multiple sequences', written in C++) [9,21] can incorporate positional information by implementing the model from the Methods section in a Gibbs sampler. A-GLAM already has several desirable features when predicting transcription factor binding sites (TFBSs). First, it optimizes motif width automatically, without user input. Second, it reports theoretically accurate E-values for candidate TFBSs. Finally, it implements a theoretically sound context-dependent Markov background model, which yielded better predictions than different, *ad hoc* Markov background models or the conventional background model of independent bases [22]. With its Markov background model, a rigorous statistical evaluation showed that even before the addition of positional information, A-GLAM's predictive accuracy was

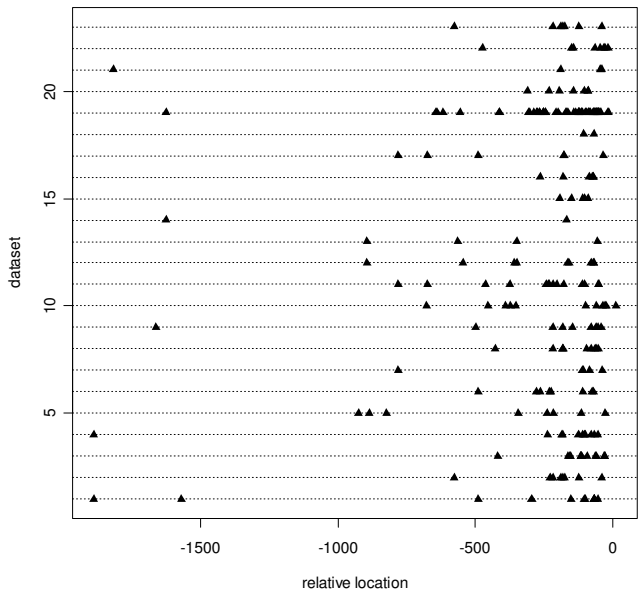

Figure 2
Distribution of known locations of binding site in TSS Tompa dataset. The x-axis is anchored on the TSS, denoted as location 0. All sequences in each test subset are collapsed into a single line; hence the 23 data subsets are shown as 23 different horizontal lines. Each data subset contains TFBSs corresponding to a single specific transcription factor.

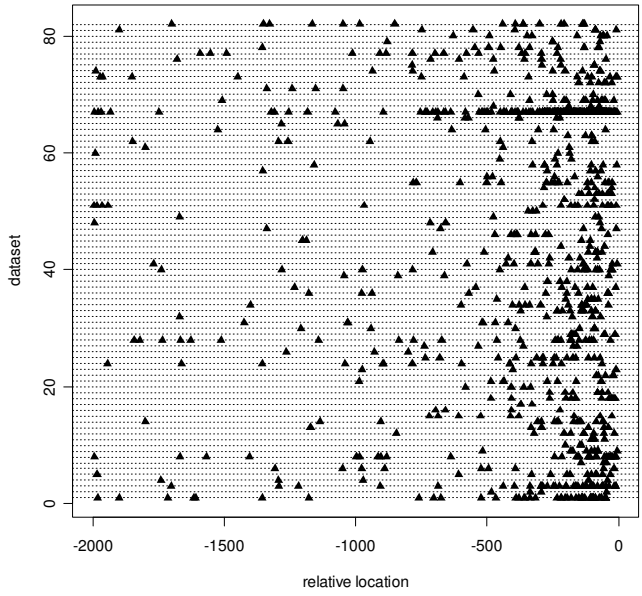

Figure 3
Distribution of known locations of binding site in TRANSFAC dataset. The x-axis is anchored on the TSS, denoted as location 0. All sequences in each test subset are collapsed into a single line; hence the 82 data subsets are shown as 82 different horizontal lines. Each data subset contains TFBSs corresponding to a single specific transcription factor.

competitive with any state-of-the-art motif-finding tool [22].

At the outset, we point out that all motif-finding tools have had notorious difficultly with the original Tompa dataset [2]. Our TSS Tompa test dataset is even more difficult than the original Tompa dataset. Its data subsets often contained fewer sequences than the corresponding original Tompa subset. Moreover, our sequences were on average longer than the corresponding original Tompa sequence. Thus, conventional motif-finding tools should perform more weakly on our TSS Tompa test dataset than on the original Tompa dataset.

The Bayesian model in this paper combines sequence and positional information to predict putative TFBSs. Its implementation in A-GLAM permits users either to accept

Table 3: The effect of truncating the sequence upstream of the TSS

Sequence range	TSS Tompa Dataset			TRANSFAC Dataset		
	Without positional info	With positional info	p-value	Without positional info	With positional info	p-value
[-2000, 0]	-0.008	0.101	0.002	-0.009	0.027	10^{-8}
[-1000, 0]	0.086	0.098	0.583	0.050	0.066	0.112
[-500, 0]	0.125	0.133	0.338	0.077	0.078	0.070
[-250, 0]	0.139	0.139	0.054	0.094	0.076	0.603

The first column shows the sequence range upstream of the TSS given as input to A-GLAM. The change of CCC from modes with and without positional information for the TSS Tompa and TRANSFAC datasets is displayed in the corresponding groups of three columns. The third column of each group shows a Wilcoxon p-value, which evaluates the difference between the CCCs in the previous two columns. Because not all TFBSs in our datasets are known, small improvements in the CCC correspond to true improvements of unknown magnitude. In particular, e.g., in the Table, two CCC values rounded to 0.139 have unseen decimals different enough to have a p-value of 0.054. To view results for individual sites in the Tompa dataset, see Supplementary Table 7 [see Additional file 1].

our default hyperparameters η for the prior distribution or to select their own. Although complete flexibility in the selection of hyperparameters can permit inappropriate or excessively aggressive choices, extensive cross-validation showed that the usual priors place mild restrictions on the predictions, so the model is very robust against misspecification of its hyperparameters or, by extension, to the locations of known sites. In other words, the prior does not dictate the alignment; instead, it loosely guides the alignment and permits the data to "speak for themselves". If motifs do not cluster by position, A-GLAM might therefore still find motifs sharing sequence but not position. We therefore make the following recommendation to users: in the absence of a strong reason to the contrary, they should accept A-GLAM's default hyperparameters.

To use positional information to find biologically active sites, A-GLAM's positional model requires the input sequences to be anchored on a genomic landmark, e.g., to find TFBSs, the model might be anchored to TSSs. Because a single gene might correspond to several alternative TSSs [23], however, TSS multiplicity might initially appear to cause problems. Moreover, the TSS itself can have either "sharp" or "broad" positional preference within a promoter [24]. Variability of the TSS position within a promoter reduces the positional information available to A-GLAM, possibly explaining the uneven improvement in prediction across our data subsets. A-GLAM's statistical model examines sequence as well as positional information, however, so it retains robustness against a mild misspecification of the TSS, say, within a few hundred bases of the true position, so alternative TSSs or TSSs with a typical broad positional distribution are unlikely to degrade predictions seriously when positional information is used. A-GLAM's users should note, however, if a TSS is specified, e.g., a kilobase away from the relevant position, positional information might severely distract A-GLAM from finding the desired TFBSs. On the other hand, however, different positions relative to the TSS containing exactly the same sequence have long been known to be associated with different TFBS biological functions [25]; in other cases, they might also be associated with alternative TSSs or TSSs with a broad positional distribution. Up to now, because computational studies of positional control of transcription have had to rely on *ad hoc* methods, A-GLAM now has a unique potential among general motif-prediction tools. Even if two functionally different sets of TFBSs have similar motifs, A-GLAM can differentiate them by position alone and report the two sets separately. It would be very interesting if someone using A-GLAM identified two sets of TFBSs of similar sequence corresponding to two different functionalities or TSSs.

The sequences in our study used the upstream positions from -2000 to 0 bp relative to the TSS to evaluate A-GLAM's accuracy in predicting TSSs. Because our purpose in this article was to evaluate A-GLAM's ability to find biologically active sequence motifs in general, there is no scientific reason not to use the 3' UTR region as a "genomic anchor" to identify nearby regulatory elements. A similar statement applies to any set of regulatory elements (e.g., TFBSs, miRNA binding sites, etc.) around any genomic landmark (e.g., the TSS, the 3' UTR, etc.).

Indeed, if its main purpose was not evaluation of the predictive accuracy of A-GLAM's positional model, this article could have restricted its input sequences to intervals downstream of the TSS, e.g., [0, 1000] bp instead. With the TSS still providing the genomic anchor, A-GLAM could have searched for motifs associated with, e.g., 5' UTRs or translation start sites, which are usually within a few hundred base pairs downstream of a TSS. Thus, positional restrictions on the input sequence could focus A-GLAM's search on sequence motifs with different biological functions.

In practice, however, restricting the input interval requires great care. Unlike the TFBSs in our test datasets, many sequence motifs have poorly characterized distributions. On one hand, excessively stringent truncation of the input interval to, say, [-125, 0] would probably have removed many TFBSs from consideration in our study. On the other hand, positional modeling generally improved the accuracy of motif prediction, never hurting it significantly, even when input sequences were truncated. In the search for novel sequence motifs, therefore, we recommend that the use of Bayesian positional modeling on an input sequence whose length is generous (but not too generous) relative to the locations of known motifs.

Since the previous study showed that A-GLAM is one of the top performers among existing tools for *de novo* TFBS discovery [22], we believe that A-GLAM now easily outperforms its competitors whenever positional information is available and relevant. "Positional genomics" exploits the information provided by genomic landmarks (like the TSS), yielding a "poor man's alignment", even when the precise sequence alignments are unavailable. Given the power of comparative genomics, which depends on accurate alignments, positional genomics presents many interesting possibilities.

Conclusion

We proposed a Bayesian model for incorporating positional preference of TFBS with respect to a genomic landmark, e.g., a TSS. The results on our test datasets show that a positional model can produce statistically significant improvements in the accuracy of motif prediction. Our cross-validation study shows that the prior distribution of our positional model is robust against mild misspecifica-

tion of its parameters. Our study of truncated input sequences indicates that the positional model provides a superior and more robust strategy than sequence truncation, especially when the positional preferences of sequence motifs are not well characterized.

Availability

The A-GLAM program and all datasets relevant to this article can be found online [26].

Project name: A-GLAM 2.1

Project home page: ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/2008-02-20/

Operating system: Linux

Programming language: C++

Licence: No license required.

Methods

The two test datasets

Our first test dataset was a subset of the "real" human sequences in the "original Tompa dataset", from [2]. The original Tompa dataset does not annotate any experimentally verified TSS positions, which were supplied from the Database of Transcription Start Sites (DBTSS) [27], as follows. BLAT [28] searched the DBTSS for hits to sequences in the original Tompa dataset. The DBTSS is incomplete, so when BLAT returned no hits in a sequence, the corresponding sequence was discarded. After the BLAT search, the dataset contained 26 data subsets, each composed of human sequences with a known TSS, and each corresponding to a single type of TFBS, like the original Tompa data subsets. We then discarded data subsets with 0 or 1 sequences, resulting in our "TSS Tompa dataset", which contained 23 data subsets. Each data subset contained from 2 to 26 sequences, and each sequence contained any number of known TFBSs, including 0. To encompass systematically all known TFBSs in the sequences, each sequence was expanded to contain proximal promoter regions from -2000 to 0 bp (upstream) relative to the corresponding TSS.

Our second test dataset was constructed from: (1) the latest human genome build (NCBI Build 36, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/); (2) transcriptional start sites (TSS) from the database of transcription start sites (DBTSS) [27]; and (3) experimentally characterized TFBSs from the TRANSFAC database (professional version 11.2) [29]. Briefly, TSS and TRANSFAC sites were mapped to the human genome using MegaBLAST [30], yielding a set of proximal promoter DNA sequences [15,31] annotated with experimentally characterized TSSs and TFBSs. In this paper, the resulting sequences are called our "TRANSFAC dataset". The TRANSFAC dataset contains 82 data subsets, each subset containing 2 to 101 sequences, and each sequence containing at least one instance of known TFBSs. Like the TSS Tompa data subsets, each data subset corresponded to a single type of TFBS. Like our TSS Tompa dataset, the range of TRANSFAC dataset is from -2000 to 0 bp (upstream) relative to the corresponding TSS.

A standard measure of prediction accuracy, the correlation coefficient, described elsewhere [22], evaluated TFBS predictions within our test dataset.

A Bayesian model for positional preferences

Our model for TFBSs uses two sources of information: sequence and position. We discuss sequence later, to focus on the novelties of position first.

Figure 2 displays the positions of all known TFBSs within the data subsets of the TSS Tompa dataset. Figure 2 collapses all sequences in each test subset into a single line anchored at the TSS. Thus, the 23 lines represent the 23 data subsets. Figure 2 shows that the TFBSs in several data subsets display positional preferences with respect to the TSS. Many TFBSs are upstream of the TSS, possibly clustered around certain positions. Accordingly, we search for TFBS positions that are normally distributed, with unknown center and dispersion, near the TSS. (Mathematical convenience facilitates the choice of the normal distribution.) Analogous to Figure 2, Figure 3 contains the positions of all known TFBSs in the TRANSFAC dataset. The TRANSFAC dataset displays the same basic distributional characteristics as the TSS Tompa dataset in Figure 2.

Fix a data subset in Figure 2 or 3, and assume it contains some number n of unknown TFBSs with locations $x_1,...,x_n$ relative to the TSS. For later reference, let $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$ and $s_n^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ be the sample mean and sample standard deviation. Assume $\mathbf{x} = (x_1,...,x_n)$ constitute independent samples from a Normal (μ, λ) distribution, with mean μ and reciprocal variance (also known as "precision") $\lambda = 1/\sigma^2$. Given the normal parameters $\theta = (\mu, \lambda)$, the positions $\mathbf{x}$ have the likelihood function

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\lambda\sum_{i=1}^{n}(x_i - \mu)^2\right\}. \quad (1)$$

Parenthetically, to avoid confusion, the sequence locations $x_1,...,x_n$ are integers, but the use of continuous distributions (e.g., the normal) as approximations simplifies

the algebra enormously. Similarly, the locations $x_1,...,x_n$ might be confined to a finite interval (e.g., they might be within a finite piece of DNA). The seemingly unrestricted normal distribution remains appropriate, however, because its rapidly vanishing squared exponential form (as in Eq) effectively confines its samples to a finite interval.

Now, let the normal parameters $\theta = (\mu, \lambda)$ have a uniform-gamma prior distribution, in which μ and λ have independent prior distributions. The prior for μ is the continuous Uniform $[a, b]$ distribution on some closed interval $[a, b]$ $(a < b)$, with constant density $p(\mu) = (b - a)^{-1}$ for $\mu \in [a, b]$. The prior for λ is a Gamma(α, β) distribution with parameters $\alpha, \beta > 0$, with density

$$p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$$

for $\lambda \geq 0$. The uniform-gamma prior distribution for $\theta = (\mu, \lambda)$ therefore has the joint density function

$$p(\boldsymbol{\theta}) = (b-a)^{-1} \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda),$$

for $\mu \in [a, b]$ and $\lambda \geq 0$.

Practical suggestions for the numerical values of α and β are given below.

Our aim is to provide a figure of merit for Gibbs sampling based on the predictive distribution $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$ of the locations $\mathbf{x}$. Gibbs sampling conditions on the locations $\mathbf{x} = (x_1,...,x_n)$ to determine the conditional predictive distribution of the location $x_{n+1} = x$ of another TFBS (see Eq (2) below). After extensive algebraic manipulation of the relevant integrals, the conditional predictive distribution is

$$p(x \mid \mathbf{x}) = \frac{p(\mathbf{x},x)}{p(\mathbf{x})} = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\Gamma\left(\frac{1}{2}\nu\right)} (\nu\pi)^{-1/2} \sigma^{-1} \left\{ 1 + \frac{(x-\bar{x}_n)^2}{\nu\sigma^2} \right\}^{-\frac{1}{2}(\nu+1)}, \quad (2)$$

a Student t-distribution whose parameters are $\nu = 2\left[\alpha + \frac{1}{2}(n-1)\right]$, $\bar{x}_n$, and

$$\sigma^2 + \frac{n+1}{n} \frac{\beta + \frac{1}{2} n s_n^2}{\alpha + \frac{1}{2}(n-1)}.$$

The t-distribution has mean $\bar{x}_n$ for $\nu > 1$ and variance $\nu(\nu-2)^{-1}\sigma^2$ for $\nu > 2$.

The result in Eq (2) ignores the restriction $\mu \in [a, b]$. If $[a, b]$ covers most of the range $[a', b']$ of the locations $\mathbf{x}$ (e.g., $a - 3\sigma < a' < b' < b + 3\sigma$), then analysis will confirm that under appropriate mathematical hypotheses, Eq (2) approximates the desired conditional predictive distribution accurately.

The prior distribution is fully specified by a list of the hyperparameters a, b, α, and β. As indicated above, any sufficiently generous interval $[a, b]$ containing the locations $\mathbf{x}$ suffices for present purposes. The input sequence range (e.g., in the case of TSS Tompa's dataset as well as TRANSFAC dataset, from -2000 to 0 bp relative to the corresponding TSS) is a practical choice for $[a, b]$. In contrast, the selection of α and β can be delicate. On one hand, a user can provide subjective preferences for α and β, yielding a precision λ with mean $\alpha\beta^{-1}$ and variance $\alpha\beta^{-2}$. On the other hand, α and β can be estimated from the distributions of experimentally verified TFBSs, as follows.

Suppose we have k data subsets, where the i-th data subset $(i = 1,...,k)$ yields a known vector $\mathbf{x}_i$ of locations for a particular TFBS. Each data subset $\mathbf{x}_i$ corresponds to a different set of hyperparameters $\{\theta_i = (\mu_i, \lambda_i)\}_{i=1,...,k}$ chosen from a common uniform-gamma prior with unknown parameters $\eta = (\alpha, \beta)$. The predictive distribution of the data is

$$p(\mathbf{x}_1, ..., \mathbf{x}_k \mid \boldsymbol{\eta}) = \left\{ \int p(\mathbf{x}_1 \mid \boldsymbol{\theta}_1) p(\boldsymbol{\theta}_1 \mid \boldsymbol{\eta}) d\boldsymbol{\theta}_1 \right\} \cdots \left\{ \int p(\mathbf{x}_k \mid \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k \mid \boldsymbol{\eta}) d\boldsymbol{\theta}_k \right\}.$$

Maximization of the predictive distribution yields the so-called type-II maximum likelihood estimate for $\eta = (\alpha, \beta)$ [32].

In this study, based on our two datasets, the type-II maximum likelihood estimate of α and β were selected. The value of α was 0.8424; of β, 25790 for TSS Tompa dataset; α, 0.5825, β, 12818, for TRANSFAC dataset. Thus, the distribution of the precision λ had mean 3.27×10^{-5} (4.54×10^{-5}, for TRANSFAC dataset), giving the scale parameter $\sigma = \lambda^{-1/2}$ an approximate mean 175 (148, for TRANSFAC dataset). (The lengths of typical input sequences are several hundreds to a couple of thousand, e.g., in our dataset,

the lengths are all 2000.) Now, 95% of the realizations from a Normal (μ, λ) distribution fall into the interval (μ-2σ, μ+2σ) of length 4σ. Because 4σ = 4(175) = 700 (592, for TRANSFAC dataset), the above selection of α and β makes the prior distribution quite broad, permitting the data "to speak for themselves".

Some comments on the distributional choices for the prior and likelihood

The normal distribution might be challenged as an inappropriate form for the likelihood. In most of the data subsets in Figure 2 or3, it is completely justifiable, but does appear untenable for a few. Although mathematical convenience facilitates the choice of a normal distribution, one could propose alternative distributional forms, usually at the expense of greater complexity. The normal distribution is quite adequate, however, when modeling any cluster lacking distant "orphan" locations.

Similarly, a uniform prior for the normal mean μ might be challenged. In fact, we implemented the same model with a normal-chi-square prior for $\theta = (\mu, \lambda)$. In our hands, both models produced comparable results on our test dataset (data not shown).

Gibbs sampling using both sequence and position

As noted above, Gibbs sampling requires only conditional predictive distributions. Because of the uniform prior for μ, multiplying the conditional predictive distribution in Eq (2) by (an ultimately irrelevant factor of) $(b - a)$ yields an approximation for the conditional predictive odds ratio with respect to the uniform background model. Taking logarithms and adding subscripts for "location", yields a log-odds score $\Delta s_{[l]}(x_{[l]} \mid \mathbf{x}_{[l]})$ for location.

Now, consider the sequence information. Let the n locations $\mathbf{x}_{[l]}$ initiate subsequences $\mathbf{x}_{[s]}$ of length w (for "window"). Let the count of nucleotide j in the i-th column of the window be $c_{i,j}$, so the total count in each position is $c = \sum_{(j)} c_{i,j} = n$. As in the conditional predictive distribution above, add another subsequence $\mathbf{x}_{[s]}$ of length w to the data. Let $\delta[i, j]$ equal 1 if the new subsequence contains nucleotide j in its i-th position, and 0 otherwise. Our previous work [9] postulated a familiar model [7,8], that the TFBS sequences follow a multinomial motif model with a Dirichlet prior. In the prior, the nucleotide pseudo-counts were $\{a_j\}$ $(a = \sum_{(j)} a_j)$. The background model was the so-called "independent letters model" with probabilities $\{p_j\}$. Effectively, our previous work gave the conditional log-odds ratio of the subsequence $x_{[s]}$, given the subsequences $\mathbf{x}_{[s]}$, as

$$\Delta s_{[s]}(x_{[s]} \mid \mathbf{x}_{[s]}) = \sum_{i=1}^{w} \sum_{j=1}^{4} \delta[i, j] \log\left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right]. \tag{3}$$

If sequence and position are independent in both the motif and background models, the corresponding conditional predictive log-odds ratio is $\Delta s(x \mid \mathbf{x}) = \Delta s_{[s]}(x_{[s]} \mid \mathbf{x}_{[s]}) + \Delta s_{[l]}(x_{[l]} \mid \mathbf{x}_{[l]})$. Conditional predictive log-odds ratios can be added to generate the log-odds ratios for any dataset $\mathbf{x}$ step by step. Thus, Eqs (2) and (3) completely specify a predictive log-odds ratios for use as the figure of merit in Gibbs sampling. The present article actually replaces the independent letters model for the sequence background with a Markov model of order 3 [22], but the principles are the same.

Having established the separate roles of sequence and location, we drop the subscripts [s] and [l] below, particularly in x_i, which now represents the sequence and location of the i-th candidate TFBS.

A p-value for each candidate TFBS

For consistency with other computer programs (and because it makes little practical difference), to calculate a p-value for the i-th candidate TFBS x_i, we consider the self-predictive score $\Delta s(x_i \mid \mathbf{x})$, where $\mathbf{x} = (x_i, \ldots, x_i, x_n)$ includes x_i. Because sequence and location are independent variates in both the motif and background models, the distribution of $\Delta s(x_i \mid \mathbf{x})$ is a convolution, i.e.,

$$\mathbb{P}\{\Delta s(x_i \mid \mathbf{x}) \geq t\} = \sum_{(r)} \mathbb{P}\{\Delta s_{[s]}(x_{i,[s]} \mid \mathbf{x}_{[s]}) = r\} \cdot \mathbb{P}\{\Delta s_{[l]}(x_{i,[l]} \mid \mathbf{x}_{[l]}) \geq t - r\}.$$

Existing methods [33,34] determine the distribution of $\Delta s_{[s]}$, and the distribution of $\Delta s_{[l]}$ is known. Thus, a p-value can be assigned to each candidate site.

k-fold cross-validation for sensitivity of hyperparameter selection

The k-fold cross-validation method estimates error rates in classification problems accurately [35]. The k-fold cross-validation splits the available data containing known classification labels into k mutually exclusive "partitions", so that each partition contains about the same amount of data. It then sets aside one of k partitions as the test set, and uses remaining k - 1 partitions as a training set to estimate the statistical parameters underlying the classification rule. After repeating the estimation process k times, leaving out each partition in turn, the average of the resulting classification errors estimates the error rate of the rule. The choice of 5 or 10 for k generally overcomes the

effects of replicated data, which would otherwise render the test and training data unduly similar [35]. In the present context, known sites provide estimates of the hyperparameters $\eta = (\alpha, \beta)$. In our study, cross-validation with $k = 5$ partitions was most appropriate to address over-fitting, because we have only 23 different datasets in TSS Tompa dataset. To illustrate the 5-fold cross-validation, consider the partition 23 = 5 + 5 + 5 + 4 + 4. First, set aside the first "5" of the 23 data subsets as the test set $\mathbf{x}_1$, and estimate the hyperparameters η by maximizing the value of $p(\mathbf{x}_2,...,\mathbf{x}_5|\eta)$, where $\mathbf{x}_2,...,\mathbf{x}_5$ are the 18 = 5 + 5 + 4 + 4 training sets. With the estimated hyperparameters η, A-GLAM then makes predictions on the test set $\mathbf{x}_1$. The 5-fold cross-validation then repeats the procedure, taking each of the partitions $\mathbf{x}_2,...,\mathbf{x}_5$ in turn as the test set.

To eliminate the results' dependence on the partition, the partition was chosen randomly 100 times, and the results averaged.

A-GLAM Settings for the Test Predictions

To compare the model with positional information and the model without positional information (i.e., using sequence alone), we ran A-GLAM in the ZOOPS (Zero or One Occurrence Per Sequence) mode, where A-GLAM reports zero or one instance of the motif element for each sequence. Somewhat arbitrarily, we restricted the search space to the strands in the test dataset, without the complementary strands.

Authors' contributions

N–KK proposed the Bayesian model, implemented it in A-GLAM, and ran the program on the test datasets; KT and LM–R generated the test datasets and extracted the transcription start site information; JLS conceived and supervised the study.

Additional material

Additional file 1

Additional alignments for the TSS Tompa dataset and the complete data corresponding to the summary in Table 3. Supplementary Tables 1–6 contain additional alignments for the TSS Tompa dataset. Supplementary Table 7 summarizes truncation effects for the TSS Tompa dataset; Supplementary Table 8, for the TRANSFAC dataset.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-262-S1.doc]

Acknowledgements

The authors thank Sergey Sheetlin for helpful discussion. This research was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

References

1. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-1419.
2. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23(1):**137-144.
3. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15(7-8):**563-577.
4. Pavesi G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes An algorithm for finding signals of unknown length in DNA sequences.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W199-203.
5. Sinha S, Tompa M: **YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic Acids Res* 2003, **31(13):**3586-3588.
6. Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization.** *Machine Learning* 1995, **21:**51-83.
7. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262(5131):**208-214.
8. Liu JS, Neuwald AF, Lawrence CE: **Bayesian models for multiple local sequence alignment and Gibbs sampling strategies.** *J Amer Statistical Assoc* 1995, **90:**1156-1169.
9. Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL: **Alignments anchored on genomic landmarks can aid in the identification of regulatory elements.** *Bioinformatics* 2005, **21:**1440-1448.
10. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **296(5):**1205-1214.
11. Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites.** *Nucleic Acids Res* 2003, **31(13):**3580-3585.
12. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29(3):**774-782.
13. Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H: **Combining frequency and positional information to predict transcription factor binding sites.** *Bioinformatics* 2001, **17(11):**1019-1026.
14. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters.** *Genome Res* 2004, **14(15628):**1562-1574.
15. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D: **Statistical analysis of over-represented words in human promoter sequences.** *Nucleic Acids Research* 2004, **32(3):**949-958.
16. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31(1):**374-378.
17. Li N, Tompa M: **Analysis of computational approaches for motif discovery.** *Algorithms Mol Biol* 2006, **1:**8.
18. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **High-resolution computational models of genome binding events.** *Nat Biotechnol* 2006, **24(8):**963-970.
19. Defrance M, Touzet H: **Predicting transcription factor binding sites using local over-representation and comparative genomics.** *BMC Bioinformatics* 2006, **7:**396.
20. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.** *Bioinformatics* 2001, **17(12):**1113-1122.

21. Frith MC, Hansen U, Spouge JL, Weng Z: **Finding functional sequence elements by multiple local alignment.** *Nucleic Acids Res* 2004, **32(1):**189-200.
22. Kim NK, Tharakaraman K, Spouge JL: **Adding sequence context to a Markov background model improves the identification of regulatory elements.** *Bioinformatics* 2006, **22(23):**2870-2875.
23. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30(1):**328-331.
24. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38(6):**626-635.
25. Ptashne M: **Lambda's switch: lessons from a module swap.** *Curr Biol* 2006, **16(12):**R459-62.
26. **John Spouge's Research Group** [http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/]
27. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006.** *Nucleic Acids Res* 2006, **34(Database issue):**D86-9.
28. Kent WJ: **BLAT - The BLAST-like alignment tool.** *Genome Res* 2002, **12(4):**656-664.
29. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34(Database issue):**D108-10.
30. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7(1-2):**203-214.
31. Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK: **Transposable elements donate lineage-specific regulatory sequences to host genomes.** *Cytogenetic and genome research* 2005, **110(1-4):**333-341.
32. Berger JO: **Statistical Decision Theory and Bayesian Analysis.** 2nd edition. New York , Springer-Verlag; 1985.
33. Huang H, Kao MC, Zhou X, Liu JS, Wong WH: **Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification.** *J Comput Biol* 2004, **11(1):**1-14.
34. Kann MG, Sheetlin SL, Park Y, Bryant SH, Spouge JL: **The identification of complete domains within protein sequences using accurate E-values for semi-global alignment.** *Nucleic Acids Res* 2007, **35(14):**4678-4685.
35. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning : data mining, inference, and prediction.** New York , Springer; 2001.