Short Segment Frequency Equalization: A Simple and Effective Alternative Treatment of Background Models in Motif Discovery

Kazuhito Shida

Institute for Material Research, 2-1-1 Katahira, Aoba-ku,
980-8577 Sendai, Japan
shida@imr.edu

Abstract. One of the most important pattern recognition problems in bioinformatics is the *de novo* motif discovery. In particular, there is a large room of improvement in motif discovery from eukaryotic genome, where the sequences have complicated background noise. The short segment frequency equalization (SSFE) is a novel treatment method to incorporate Markov background models into *de novo* motif discovery algorithms, namely Gibbs sampling. Despite its apparent simplicity, SSFE shows a large performance improvement over the current method (***Q/P*** scheme) when tested on artificial DNA datasets with Markov background of human and mouse. Furthermore, SSFE shows a better performance than other methods including much more complicated and sophisticated method, Weeder 1.3, when tested with several biological datasets from human promoters.

Keywords: Motif discovery, Markov background model, Eukaryotic promoters, Stochastic method, Gibbs sampling.

1 Introduction

Reliable *de novo* motif discovery remains an important problem of pattern recognition that remains unsolved by bioinformatics[1-7], in particular, when subjects are transcription factor binding sites (TFBS) in eukaryotic genomes[8] such as those of fruit fly, mouse, and human: Eukaryotic sequences tend to have a more complicated statistical structure[9] than prokaryotic sequences do. Assuming that the input sequence is a mixture of two sequences generated from two statistical information sources, the Markov background model (noise) and the motif model (signal), many motif discovery algorithms seek a maximally differentiated motif model[3, 10-13] from the given background. It is understandable that the separation of signal and noise is difficult when the noise has complicated statistical structures.

However, these two information sources have a difference in their *spatial scale*, too. In many cases considered, the motif width is greater than that of the order of the Markov background model. Although a weak long-range correlation is reported on genomic sequence, the magnitude of the correlation is a decreasing function of correlation length[14, 15]. Therefore, it is clear that the non-motif information, the "noise" for motif discovery algorithms, is concentrated in the short-range regime of background

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 354–364, 2009.

statistics. It is possible to suppress the noise and enhance the performance of the motif discovery algorithms by selectively reducing the magnitude of the short-range or high-frequency portion of the sequence information. In other words, we need a sort of "low-pass filter" of the sequence information.

A "high-pass filter" of sequence information is already realized and used to evaluate the statistical significance of alignments[10, 16]. For example, if the input sequences are cut into numerous non-overlapping pieces of length x and re-organized in a randomly shuffled order, all information contained in the spatial scale of x+1 or longer will be randomized and erased, without imparting a large effect on information found in shorter scales in input sequences. This is exactly why the shuffled sequence is useful as the null hypothesis of sequence alignment and motif discovery.

A sequential "low-pass filter" based on the shuffling principle seems to be difficult to realize. This report presents the proposal of adding a very simple modification, a "built-in filter", to the conventional Gibbs sampling, thereby rendering the resultant sampling behavior as low-pass filtered and noise-tolerant. This filtering method is called short segment frequency equalization (SSFE).

2 Method

2.1 Conventional Method

We take Gibbs sampling[11] as our starting point: Gibbs sampling is a type of Markov Chain Monte Carlo (MCMC) method that samples all possible blocks (gapless alignments) with width w in N input sequences with length L, at a probability linear to

$$\boldsymbol{Q/P},\quad \boldsymbol{Q}=\prod_{y=1}^{N} Q(row_y),\ \ \boldsymbol{P}=\prod_{y=1}^{N} P(row_y), \tag{1}$$

where $Q(row_y)$ and $P(row_y)$ respectively signify the likelihood of the y-th row of the block in the current motif model and the given background model. The likelihoods assigned to the entire block are denoted in boldface. Usually the motif model is a position weight matrix (PWM) from which likelihood Q is calculated. The value of a PWM element, q_{si}, is the number ratio of letter s in the i-th column of the block, calculated with an appropriate pseudocount.

The following is the outline of Gibbs sampling based motif discovery with conventional treatment of a Bernoulli background. First, the current PWM, q, is calculated from the current alignment. In the row update (row resampling) phase of conventional Gibbs sampling, the y-th row of the current alignment is updated to be one of all possible length w substrings (segments) in the y-th input sequence sampled with probability

$$Q_x / P_x\,, \tag{2}$$

where Q_x signifies the likelihood that the x-th substring (comprising the x-th to x+w-1 -th letters of the sequence, $s(x)$ ~ $s(x+w\text{-}1)$) comes from the current model denoted by q; P_x is the likelihood that the same substring comes from a Bernoulli background denoted by p,

$$Q_x=\prod_{i=0}^{w-1} q_{s(x+i)\,i}\,,\quad P_x=\prod_{i=0}^{w-1} p_{s(x+i)}\,. \tag{3}$$

After the update is done, y is changed to $(y+1)mod(N)$, such that all rows are updated in a cyclic manner. Subsequently, the entire process is repeated starting from the updated alignment.

When Markov background models are used, the P part of the transition probability that is used in the original Gibbs sampling is changed to a sequence-dependent one, as

$$P_x = P(s(x+0)s(x+1)s(x+2)\dots s(x+w-1)). \tag{4}$$

The value of P_x is given by the following formula when w is greater than the order of Markov background model, m,

$$P_x = P(s(x+0)\dots s(x+m)) \prod_{i=1}^{w-(m+1)} \frac{P(s(x+i)\dots s(x+i+m))}{\sum_{t=\mathrm{G,A,C,T}} P(s(x+i)\dots s(x+i+m-1)t)} \tag{5}$$

Because of its popularity[3, 17, 18], this method of incorporating background models will be designated as the "conventional" method or simply the "***Q/P***-scheme" throughout this report. The "***Q/P***-scheme" is surely an effective noise reduction scheme because it penalizes frequent m-mers to be sampled as motif. However, it should be noted that there is no mathematical proof on the quantitative correctness of the penalty.

2.2 Proposal of a New Background Treatment

Basically, SSFE method differs from the conventional Gibbs sampling scheme only in a very small but crucial point (Fig. 1): a likelihood according to "modified background model", ***P'***, is used in place of ***P***. The main characteristic of ***P'*** is that the behavior of the "***Q/P'***-scheme" is almost totally unbiased toward any short segment.

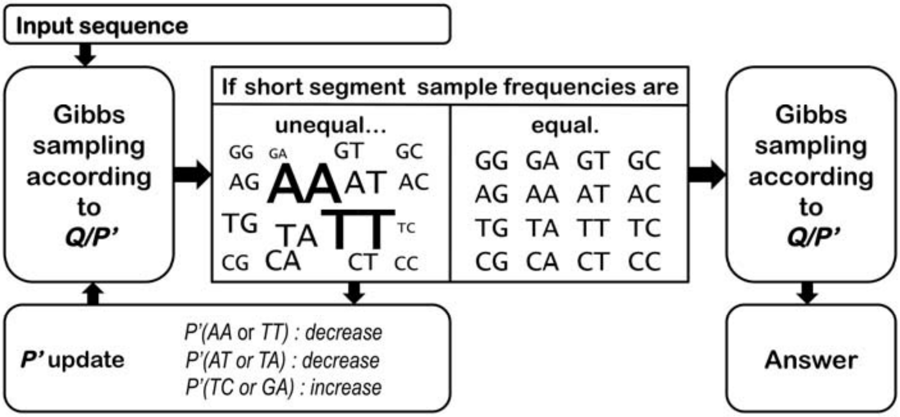


Fig. 1. Schematic explanation of the SSFE scheme ($m=1$). The size of the letters indicates the frequency of dimers. The equalization stage (left half) iteratively adjusts the background parameters such that no $m+1$-mer is preferentially sampled in the detection stage (right half). Note that each stage is a simple Gibbs sampling by itself.

Actually, ***P'*** is obtainable with ease using the following simple iterative process. Identically to ***P***, ***P'*** is based on a Markov model, but with smaller order (therefore, "short segment"). For this report, the order of the model is chosen as *m*=2 (apparently, this is shorter than most of nucleotide motifs). First, some plausible Markov background is prepared as an initial point to start the *equalization* stage. A short Gibbs sampling is performed using a conventional scheme using ***Q/P'*** calculated from current background model. After each row updating, the newly selected length *w* segment is decomposed to *w-m* short segments (for example, "TATCGT" will be decomposed into TAT, ATC, TCG and CGT) to evaluate the frequency of *m+1*-mers to be *sampled* under current ***P'***. If the evaluated sample frequency is biased beyond an appropriate threshold, the background model is adjusted to counterbalance the bias by increasing or decreasing the background parameters by a fixed step (More sophisticated optimization methods, e.g. high-dimensional Newton–Raphson method on some "flatness of sampling" function, can actually be problematic for SSFE because it is difficult to calculate the Jacobian matrix of such goal function). Then the updated background is used to calculate ***P'*** in the next short Gibbs sampling.

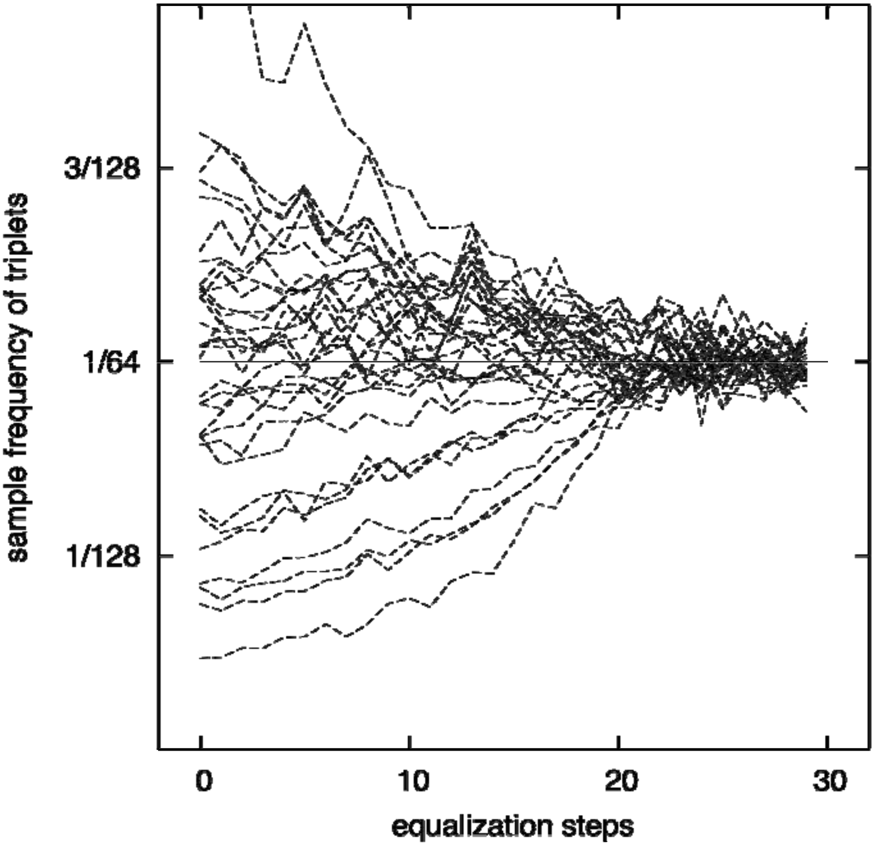


Fig. 2. Sample frequencies (Y-axis) of different triplets converge to the near equal values, 1/64, as the iteration number (X-axis) of short Gibbs sampling in the equalization stage increases

Typically, after 30–60 short Gibbs samplings, each of which is 2000–4000 steps long, the bias in the frequencies of *m+1*-mers is reduced within the threshold (see Fig. 2), which means that the background model is converged to the optimal one for balancing *m+1*-mers. From this point on, we can start the *detection* stage in which the Gibbs sampler based on the ***Q/P'***-scheme can sample any segment with any length at near-equal frequency, unless the sampling is disturbed by information retained in larger spatial scale in the input sequences. The most likely cause of such disturbance is the over-representation of mutually similar sequences with their length greater than *m+1*. In other words, it is highly probable that the disturbance is related to the biological motifs.

The selection of *m*=2 has no implication on the legitimacy of a background model with higher order. The main reason to use *m*=2 segments as the target of equalization is that, although the direct equalization of a longer segment is theoretically possible, it requires many more sampling steps in iterative adjustment (if *m*=7 is used, at least 65,536 steps are necessary for each iteration) and the sampling error will be much larger.

3 Results

An SSFE sampler is implemented in C++ language, as an extension of a previously reported motif discovery tool, GibbsST[19]. With minimum change (the order of Markov background model is changed, and the equalization stage is omitted), this SSFE sampler can precisely simulate a ***Q/P***-sampler with Markov background with any order.

The motif score must be chosen carefully: In Gibbs-sampling-based motif discovery, the value of ***Q/P*** is frequently regarded as the score and used for selection of the motif candidates. Because this is the first time that SSFE is proposed and tested, it should be tested in conditions closely resembling those of typical usage of the conventional method. Therefore, ***Q/P'*** and ***Q/P*** are used, respectively, as the score function in the test for SSFE and the conventional sampler. Both algorithms start the sampling from a number (50) of randomly generated PWM and output the best motif model with largest observed ***Q/P'*** or ***Q/P***. In short, we use a typical likelihood ratio score, but the background model might be adjusted by SSFE.

Test datasets are prepared under the following specifications.

(1) Artificial motifs implanted in a biological background. Randomly generated (*w*,*d*) artificial motif sequences are implanted randomly into artificial sequences with biologically correct statistical features: background generation is performed according to the parameters of an seventh-order Markov background model given as a part of the Weeder 1.3 toolkit for fruit fly, mouse, and human. The motif width, *w*, is set to 8 (corresponds to the order of background model). The number of mismatches per occurrence, *d*, is adjusted in conjunction with the number of sequences, *N*, and the length of an input, *L*, such that conventional Gibbs sampling shows modest success on the dataset because motifs that are too easy or too difficult to find are unable to prove differences between the two methods. The condition that is finally used is *L*=600, *N*=12, *d*=10/12. Although this condition seems slightly easier than the artificial motifs reported to be at the limit of detection possibility[4], this difference can be explained by the severe disturbance from the eukaryotic background models.

(2) Biological dataset from eukaryotic genome. Confirmed human TFBS and their flanking promoter sequences were obtained mainly from a curated database of eukaryotic promoters, ABS[20]. The TFBS with too few examples, large gaps, and overly variable structures were omitted by manual inspection to realize a modest level of difficulty. Finally, five human TF (CREB, SRF, TBP, USF, and E2F1) were used to construct our biological datasets. All sequences in the datasets have at least one TFBS (OOPS occurrence model), and the average sequence length was 504.6. In a sense, these data can be regarded as the test of SSFE for smaller N and larger w (up to 10). In addition to SSFE and conventional Gibbs sampling, two most successful motif discovery softwares— MEME (v4.1)[13] and Weeder (v1.3)[21]—are tested on this dataset. The seventh-order Markov model for human background sequence is used with Weeder, MEME, and conventional Gibbs sampling. The values of w given to these algorithms are the biologically correct ones, with one exception (TBP is processed by Weeder with w=6, because Weeder cannot use odd values of w).

The performance is evaluated as the performance coefficient S, which is defined as

$$\begin{aligned} s_i &= \max(0, \min(x_i + w, y_i + z_i) - \max(x_i, y_i)), \\ S &= \sum_{i=1}^{N} s_i / \sum_{i=1}^{N} w + z_i - s_i \end{aligned} \tag{6}$$

where x_i, y_i, and z_i respectively signify the resultant motif starting points, correct motif starting points, and correct motif width in the input sequences. This coefficient is basically the fraction of correctly discovered motif sites (1.0 represents the best performance).

In both artificial and biological tests, the "correct" length of the motif sequence is given to algorithms. Moreover, the possibilities of motif sequences in reverse strands are excluded and algorithms are not searching in reverse strands. These settings are intended to give the whole test appropriate difficulty, and not to favor SSFE sampling.

In Fig. 3, the performance observed for artificial datasets with a biological background model is shown. In the figure, the X-axis corresponds to individual datasets and the Y-axis shows performance coefficients obtained from two background cancellation methods, SSFE and Q/P, shown by upward and downward triangles, respectively. Wherever SSFE outperforms the conventional method, the gap separating two performance coefficients is shaded dark gray. Otherwise, the gap is shaded light gray. The X-axis is sorted to gather light and dark gray regions as much as possible. The average values of performance coefficients over different datasets are also portrayed in the graph. As indicated by the large dark grey areas in Fig. 3, SSFE shows marked performance enhancement over the conventional method for artificial datasets generated from human and mouse background models, although little or no improvement is apparent in the case of the fruit fly. Considering the simplicity of SSFE, this magnitude of improvement (more than two-fold increase for two of the most complicated eukaryotic background models) is surprising.

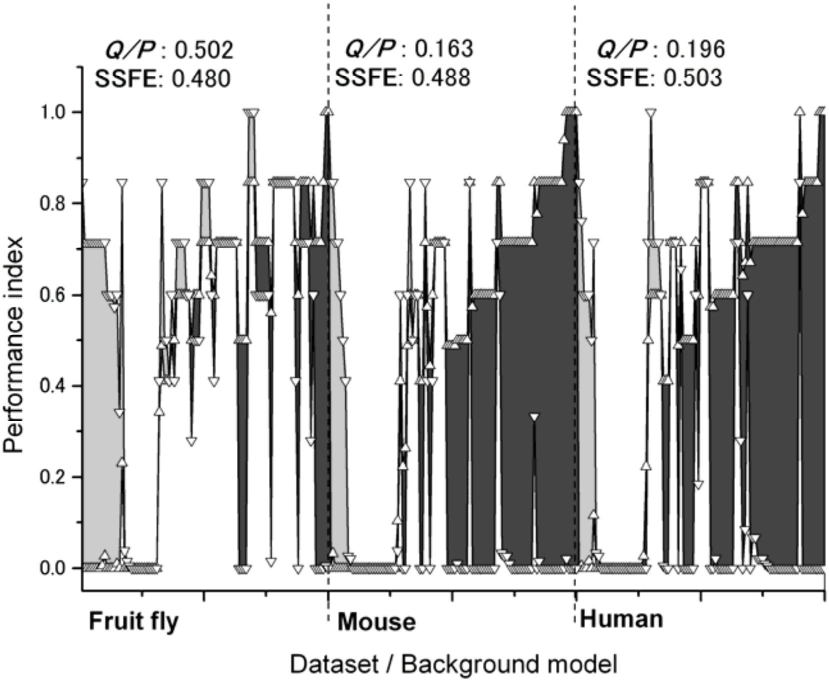
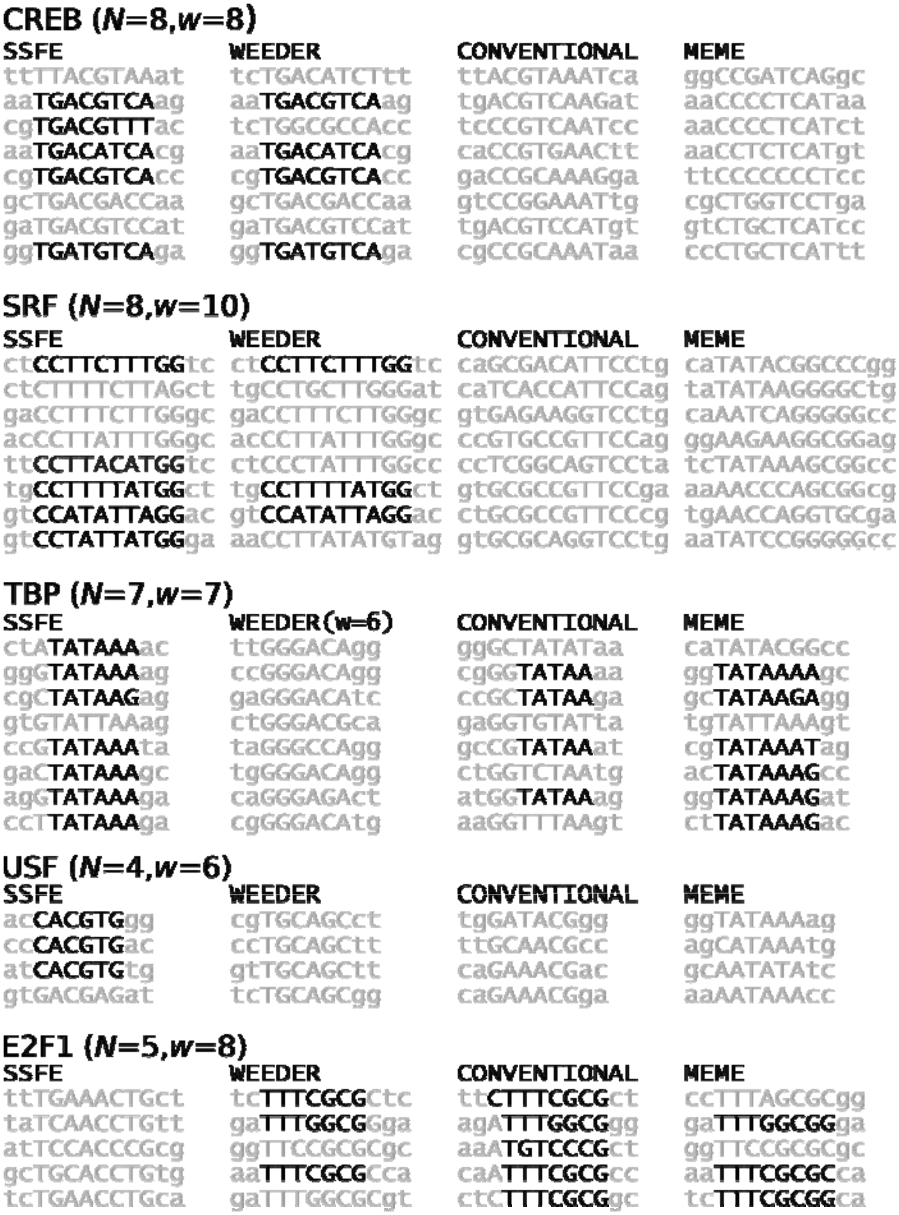


Fig. 3. Performance improvement of the SSFE scheme on artificial datasets. Percentages (Y-axis) of found sites of artificial (*w*,*d*) motifs by conventional and SSFE schemes for 300 test datasets (X-axis). The dark gray region represents that SSFE is superior to the conventional scheme; the light gray region is otherwise. Although the datasets are artificial, their background statistics are those of a fruit fly, mouse, and human.

Two features in these data might be useful to elucidate the difference between the conventional method and SSFE. First, SSFE does not increase the performance uniformly; it performs excellently for outnumbering datasets for which the conventional method performs poorly, and vice versa. Second, that the performance of SSFE is merely comparable to the conventional method in the case of fruit fly is unlikely to be a coincidence because the 8-mer distribution of the fruit fly is the least heterogeneous one among the three background models tested. The ratio of largest to smallest 8-mer frequencies is 904.3 for fruit fly, that is a much smaller value compared to human (5879.0) and mouse (12992.5). Probably, the strength of SSFE cannot be exhibited when the background model approximates a Bernoulli model. The limited performance of motif discovery based on conventional Gibbs sampling for the human and mouse background strongly suggests the incapability of conventional method to handle heterogeneous background properly. Although the data are not shown, these general trends are not changed when several other settings of *l, d, N*, are tested.

In Fig. 4, the performance observed for human promoter datasets is shown. In short, the effectiveness of SSFE is not limited to artificial motifs. In all datasets except for E2F1, the solution from SSFE is in better quality than the solution from other methods tested. It is noteworthy that E2F1 dataset also requires the largest number



CREB (N=8,w=8)

SSFE	WEEDER	CONVENTIONAL	MEME
ttTTACGTAAat	tcTGACATCTtt	ttACGTAAATca	ggCCGATCAGgc
aa**TGACGTCA**ag	aa**TGACGTCA**ag	tgACGTCAAGat	aaCCCCTCATaa
cg**TGACGTTT**ac	tcTGGCGCCAcc	tcCCGTCAATcc	aaCCCCTCATct
aa**TGACATCA**cg	aa**TGACATCA**cg	caCCGTGAACtt	aaCCTCTCATgt
cg**TGACGTCA**cc	cg**TGACGTCA**cc	gaCCGCAAAGga	ttCCCCCCCTcc
gcTGACGACCaa	gcTGACGACCaa	gtCCGGAAATtg	cgCTGGTCCTga
gaTGACGTCCat	gaTGACGTCCat	tgACGTCCATgt	gtCTGCTCATcc
gg**TGATGTCA**ga	gg**TGATGTCA**ga	cgCCGCAAATaa	ccCTGCTCATtt

SRF (N=8,w=10)

SSFE	WEEDER	CONVENTIONAL	MEME
ct**CCTTCTTTGG**tc	ct**CCTTCTTTGG**tc	caGCGACATTCCtg	caTATACGGCCCgg
ctCTTTTCTTAGct	tgCCTGCTTGGGat	caTCACCATTCCag	taTATAAGGGGCtg
gaCCTTTCTTGGgc	gaCCTTTCTTGGgc	gtGAGAAGGTCCtg	caAATCAGGGGGcc
acCCTTATTTGGgc	acCCTTATTTGGgc	ccGTGCCGTTCCag	ggAAGAAGGCGGag
tt**CCTTACATGG**tc	ctCCCTATTTGGcc	ccTCGGCAGTCCta	tcTATAAAGCGGcc
tg**CCTTTTATGG**ct	tg**CCTTTTATGG**ct	gtGCGCCGTTCCga	aaAACCCAGCGGcg
gt**CCATATTAGG**ac	gt**CCATATTAGG**ac	ctGCGCCGTTCCcg	tgAACCAGGTGCga
gt**CCTATTATGG**ga	aaCCTTATATGTag	gtGCGCAGGTCCtg	aaTATCCGGGGGcc

TBP (N=7,w=7)

SSFE	WEEDER(w=6)	CONVENTIONAL	MEME
ctA**TATAAA**ac	ttGGGACAgg	ggGCTATATaa	caTATACGGcc
ggG**TATAAA**ag	ccGGGACAgg	cgGG**TATAA**aa	gg**TATAAAA**gc
cgC**TATAAG**ag	gaGGGACAtc	ccGC**TATAA**ga	gc**TATAAGA**gg
gtGTATTAAag	ctGGGACGca	gaGGTGTATta	tgTATTAAAgt
ccG**TATAAA**ta	taGGGCCAgg	gcCG**TATAA**at	cg**TATAAAT**ag
gaC**TATAAA**gc	tgGGGACAgg	ctGGTCTAAtg	ac**TATAAAG**cc
agG**TATAAA**ga	caGGGAGAct	atGG**TATAA**ag	gg**TATAAAG**at
ccT**TATAAA**ga	cgGGGACAtg	aaGGTTTAAgt	ct**TATAAAG**ac

USF (N=4,w=6)

SSFE	WEEDER	CONVENTIONAL	MEME
ac**CACGTG**gg	cgTGCAGCct	tgGATACGgg	ggTATAAAag
cc**CACGTG**ac	ccTGCAGCtt	ttGCAACGcc	agCATAAAtg
at**CACGTG**tg	gtTGCAGCtt	caGAAACGac	gcAATATAtc
gtGACGAGat	tcTGCAGCgg	caGAAACGga	aaAATAAAcc

E2F1 (N=5,w=8)

SSFE	WEEDER	CONVENTIONAL	MEME
ttTGAAACTGct	tc**TTTCGCG**Ctc	tt**CTTTCGCG**ct	ccTTTAGCGCgg
taTCAACCTGtt	ga**TTTGGCG**Gga	agA**TTTGGCG**gg	ga**TTTGGCGG**ga
atTCCACCCGcg	ggTTCCGCGCgc	aaA**TGTCCCG**ct	ggTTCCGCGCgc
gcTGCACCTGtg	aa**TTTCGCG**Cca	caA**TTTCGCG**cc	aa**TTTCGCGC**ca
tcTGAACCTGca	gaTTTGGCGCgt	ctC**TTTCGCG**gc	tc**TTTCGCGG**ca

Fig. 4. Result of the SSFE scheme on biological data (human TFBS) compared to other algorithms. Successfully identified portions of TFBS by respective methods are marked black.

(ca. 100) of equalization steps for SSFE to converge. While MEME shows a very good performance for TBP and E2F1 and a complete failure for other datasets, Weeder shows relatively good performance for CREB, SRF, and E2F1. For TBP and USF, however, Weeder fails to present the correct answer as the most likely answer. Apparently, the performance of SSFE on these dataset is better than those of other methods tested. Considering the simplicity of SSFE, this level of performance enhancement observed for human data is remarkable.

4 Discussion

If the conventional scheme cannot handle heterogeneous backgrounds ideally, how did it manage to increase performance[3, 17, 18] in the previous reports? The answer to this question is speculated to be short low-complexity sequences in input. According to the Weeder 1.3 frequency files, the 8-mers with largest *P* in the human genome are “AAAAAAAA” and “TGTGTGTG”; for mouse, they are followed by “CACACACA” and “AGAGAGAG”. It is often pointed out[22] that these short repeats have very strong disturbance effects over Gibbs sampling. Consequently, it is plausible that the ***Q/P***-scheme was successful at least to alleviate these largest sources of problems (by imposing the maximum penalty on them) and to outperform older Gibbs sampling[11] that assumes only Bernoulli background.

The next question is more important but more difficult to answer: how it is possible for something as simple as SSFE to mark such a large increase in performance? To answer this question, more elaborate tests using a wider variety of test data should be conducted. In addition, we must develop at least some theory, not an analogy like the “low-pass filter”, on what the new score ***P’*** actually represents. The author is currently investigating the following hypothesis as a candidate of such a theory. There should be a quantitatively correct system of penalty on score, under which the bias from background model has absolutely no effect on the result of motif discovery algorithms. In SSFE, ***P’*** may work as a crude approximation of such an ideal penalty, because the equalization stage of SSFE is basically adjusting its own sampling behavior as unaffected as possible by the input sequence constituted of large amount of background and only small fraction of (often diverged) motif sequences. If this hypothesis is correct, SSFE tends to wrongfully exclude correct answers when motif sequences have a particularly large presence in the input (that is, when the motif is “easy” to discover): a good explanation for the result of SSFE on the E2F1 dataset. A possible solution for this weakness of SSFE is taking the motif score into account in the equalization stage, such that the answers with statistically meaningful level of scores will not penalized by ***P’***.

No matter what is the true strength of SSFE, its success strongly suggests a large gap in our current understanding of pattern discovery under a highly heterogeneous background model. At least, using the ***Q/P***-scheme under a heterogeneous and complicated background model must be seriously re-considered. The basic idea of SSFE can be applied to other problems in bioinformatics. The idea of equalization in terms of sub-pattern sample frequency is applicable to other motif score functions and even to sequence analyses of other types that are strongly affected by the sequence background model. The background sequence statistics is apparently a major source of inherent complexity of the biological data. Therefore, improved treatments of background models should take a much higher position in future bioinformatics for a better processing of the biological patterns. It is hoped that SSFE can serve as a good starting point for efforts in this direction.

Acknowledgments. The initial stage of this work was supported by "Special Coordination Funds for Promoting Science and Technology" of the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Reddy, T.E., DeLisi, C., Shakhnovich, B.E.: Binding site graphs: A new graph theoretical framework for prediction of transcription factor binding sites. Plos Computational Biology 3, 844–854 (2007)
2. Mahony, S., Hendrix, D., Golden, A., Smith, T.J., Rokhsar, D.S.: Transcription factor binding site identification using the self-organizing map. Bioinformatics 21, 1807–1814 (2005)
3. Liu, X.S., Brutlag, D.L., Liu, J.S.: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. 20, 835–839 (2002)
4. Pevzner, P.A., Sze, S.H.: Combinatorial approaches to finding subtle signals in DNA sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8, 269–278 (2000)
5. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14, 55–67 (1998)
6. Sinha, S., Tompa, M.: YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Research 31, 3586–3588 (2003)
7. Pavesi, G., Zambelli, F., Pesole, G.: WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. BMC Bioinformatics 8 (2007)
8. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., 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. 23, 137–144 (2005)
9. Csuros, M., Noe, L., Kucherov, G.: Reconsidering the significance of genomic word frequencies. Trends in Genetics 23, 543–546 (2007)
10. Neuwald, A.F., Liu, J.S., Lawrence, C.E.: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. 4, 1618–1632 (1995)
11. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262, 208–214 (1993)
12. Frith, M.C., Hansen, U., Spouge, J.L., Weng, Z.: Finding functional sequence elements by multiple local alignment. Nucleic Acids Res. 32, 189–200 (2004)
13. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36 (1994)
14. Messer, P.W., Bundschuh, R., Vingron, M., Arndt, P.F.: Effects of long-range correlations in DNA on sequence alignment score statistics. Journal of Computational Biology 14, 655–668 (2007)

15. Herzel, H., Trifonov, E.N., Weiss, O., Grosse, I.: Interpreting correlations in biosequences. Physica A 249, 449–459 (1998)
16. Fitch, W.M.: Random Sequences. Journal of Molecular Biology 163, 171–176 (1983)
17. 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 17, 1113–1122 (2001)
18. Narasimhan, C., LoCascio, P., Uberbacher, E.: Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. Bioinformatics 19, 1952–1963 (2003)
19. Shida, K.: GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima. BMC Bioinformatics 7 (2006)
20. Blanco, E., Farre, D., Alba, M.M., Messeguer, X., Guigo, R.: ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. Nucleic Acids Res. 34, D63–D67 (2006)
21. 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. Nucleic Acids Research 32, W199–W203 (2004)
22. van Helden, J.: The analysis of regulatory sequences. In: Chatenay, D., Cocco, S., Monasson, R., Thieffry, D., Dailbard, J. (eds.) Multiple aspects of DNA and RNA from biophysics to bioinformatics, pp. 271–304. Elsevier, Amsterdam (2005)