A Combined Approach to the Identification of Transcription Factor Binding Sites in Prokaryotes

H. K. Dai and Liang Zhao

Computer Science Department, Oklahoma State University, Stillwater, Oklahoma 74078, U. S. A.

Abstract

We present a combined approach, Gibbs sampler - Pattern Counting (GS-PC), to the identification of transcription factor binding sites (TFBSs) in prokaryotes. GS-PC first utilizes an alignment-based approach, the Gibbs sampler (the 1995 version), to derive a consensus pattern for the TFBSs present in a set of co-regulated promoter sequences. Using this consensus pattern as a drastically reduced pattern space, all obtainable 6-patterns are then enumerated, allowing up to one mismatch. After filtering off the random pattern matches, i.e., those with information content less than a user-specified cutoff value, the site alignment of each 6-pattern is evaluated for its expected frequency. All 6-patterns are thus examined, and finally the statistically most significant 6-pattern is reported, together with its site alignment. When run on three sequence sets of *Escherichia coli*, GS-PC performs comparably to the latest Gibbs sampler.

Keywords: combined approach, pattern counting, transcription factor binding sites, prokaryotes, Gibbs sampler, *Escherichia coli*.

1. Introduction

1.1. Enumerative approaches and TFBSs in prokaryotes

Traditionally enumerative approaches are mostly used in the identification of transcription factor binding sites (TFBSs) in enkarytotes, such as the yeast *Saccharomyces cerevisiae* (van helden *et al.*, 1998, 2000; Sinha and Tompa, 2002), where short, highly conserved sites are typical. For prokaryotes, such as *Escherichia coli*, the problem is usually approached via alignment-based methods, for example, the Gibbs sampler (Lawrence *et al.*, 1993; Neuwald *et al.*, 1995), MEME (Bailey and Elkan, 1995), and CONSENSUS (Hertz and Stormo, 1999). The reason is that TFBSs in prokaryotes are usually much longer than their counterparts in enkaryotes, typically 20 base-pairs long (Schneider *et al.*, 1986). A simple calculation makes it clear why the usual from-the-scratch, brute force enumerative approach is not applicable: To somewhat guarantee that the search is exhaustive, one has to inspect a pattern space of size $C(25, 6)\times 4^6 \approx 1\times 10^9$, assuming only 6-patterns are to be examined. Even for a moderate dataset, the time requirement would be prohibitive.

Various techniques have been devised to reduce the candidate pattern space, with various successes (Neuwald and Green, 1994; Sagot and Viari, 1996; Jonassen *et al.*, 1995, 1997).

1.2. Combined approaches in the identification of TFBSs

An obvious way of a combined approach is to first use an enumerative approach to perform a complete scan of all statistically significant patterns in the sequence set, and then, based on the located significant patterns, trigger an alignment process.

Implementation of this approach is first found in the late 1980's (Staden, 1989; Smith *et al.*, 1990). After an alignment is obtained, Staden (1989) further calculates its information content to obtain a ranking of those discovered patterns in terms of their statistical significances, while Smith *et al.* (1990) extract a diagnostic pattern from the alignment, which can be used to determine if a new sequence belongs to the family represented by the pattern or not.

Wolfertstetter *et al.* (1996) describe a similar approach as that of Smith *et al.* (1990). They first identify patterns of a certain length that occur in a minimum percentage of all the sequences with certain mismatches. Based on the hypothesis that in contrast to true sites, random patterns have no preferred mismatch positions, they thus can be eliminated. The remaining patterns are then extended laterally by incorporating the flanking conserved regions.

1.3. GS-PC: A combined approach to the identification of TFBSs in prokaryotes

We present here a novel combined approach to addressing this problem: Gibbs Sampler - Pattern

Counting (GS-PC). Specifically, we develop a combined approach that first utilizes a reliable alignment-based algorithm, e.g., the Gibbs sampler (Lawrence *et al.*, 93; Neuwald *et al.*, 95), to locate some strong sites from a given dataset. These sites are most likely a non-exhaustive representation of all true binding sites present in the data; they however, can be used to infer a reliable consensus sequence. This consensus sequence can then be used as a significantly reduced pattern space, and thus drastically reduce the time required to perform the pattern search. For example, for the same site length of 25 base-pairs, we now need only to check $C(25, 6) \approx 2\times10^5$ patterns. When coupled further with the powerful, innovative "fragmentation" technique which is an integral part of the Gibbs sampler, we can further reduce the pattern space to mere thousands, e.g., $C(15, 6) \approx 5\times10^3$ if "fragmentation" is set to 15 in the Gibbs sampler.

The expected frequency measure (Hertz and Stormo, 1999) is an effective metric for selecting the most statistically significant patterns, and we use it here in our algorithm.

2. Materials and Methods

2.1. Materials

A sequence dataset for the CRP protein (hereafter referred to as the Stormo CRP sequence dataset) has been successfully utilized to test several popular algorithms (Stormo and Hartzell, 1989; Lawrence and Reilly, 1990; Hertz and Stormo, 1999), and is used here as the training set for our GS-PC algorithm.

Two *E. coli* sequence datasets of LexA and purR are then used to evaluate the performance of our algorithm. They are complied as follows: The *E. coli* genes that are regulated by LexA and purR, respectively, are determined according to the compilations in McCue *et al.* (2002). The corresponding two sets of DNA upstream promoter sequences are then collected from the *E. coli* whole genome data archived in the GenBank database, accession No. NC_000913. Each promoter sequence corresponds to the whole span between the transcription start of the regulated gene and the end of next upstream gene.

2.2. Algorithm summary

The Gibbs Site Sampler (the 1995 version), with "fragmentation" enabled, is first run on the given dataset to arrive at an initial site alignment. Each site segment in this alignment is then filtered by comparing its information content (I) to the highest information content (HI) of all the site segments; those with $\log(I)/\log(HI) < cutoff$ is removed from the alignment. Here *cutoff* is a user-defined heuristic cutoff value; site segments with information content higher than this value are regarded as true binding sites, and survive this filtering process, while those with lower information content than it are considered random background matches, and are weeded out.

After the whole site alignment is thus examined, the sites left in the alignment are used to derive a simplified consensus sequence. A simplified consensus sequence here means one that has a nucleotide symbol, *a*, *c*, *g*, or *t*, at a "fragmentation turned on" position and a wildcard symbol at all other positions. All 6-patterns obtainable from this consensus constitute the candidate pattern space. Then one by one, the 6-patterns are enumerated, allowing up to one mismatch. Overlapping sites are not allowed, according to the biological motif model (Hertz and Stormo, 1999). For a 6-pattern, all sites matched are aligned together to give an alignment, and the information content of each site segment is calculated based on the whole alignment excluding itself. Then the filtering process is performed again following exactly the same procedure: each site segment with information content ratio less than the user-specified cutoff is removed from the alignment. Finally the expected frequency of the site alignment formed by all the remaining sites is calculated.

The whole process is repeated for each 6-pattern obtainable from the simplified consensus pattern. Finally, the 6-pattern with the lowest expected frequency is reported, together with the corresponding site alignment.

3. Results and Discussions

The program implemented as described in Section 2.2 is run on the Stormo CRP sequence set. The *cutoff* at both filtering points are varied between 0.10 ~ 0.70, and the *fragmentation* in the Gibbs site sampler is varied at 9, 12, 15, and 18 columns at a time. For each combination of *cutoff* and *fragmentation* settings, the GS-PC is run 10 times, and the most statistically significant 6-pattern, as measured by the least expected frequency, is reported, together with other related information, such as the site alignment, its expected frequency, and its number of repeats over the 10 runs. Table 1 summarizes the best result at each combination of the *cutoff* and *fragmentation* values.

It is obvious from Table 1 that the parameter settings with *cutoff* = 0.20 or 0.30, *fragmentation* = 15 columns give the best results, each locating 21 binding

sites, none being false positive. More impressively, the same best results are repeated six times out of 10 runs. This suggests that the proposed combined algorithm is pretty steady.

The next best results are the parameter settings with *cutoff* = 0.40 ~ 0.60, and *fragmentation* = 15 columns (again). These settings each give 19 sites, none being false positive. Again, the results are pretty steady; each alignment repeats itself at least six times out of 10 runs. It is not surprising that less sites are found with these parameter settings, since as the cutoff value is increased from 0.20 ~ 0.30 to 0.40 ~ 0.60, some true but weak binding sites are filtered off.

Also note that in all these five cases, *fragmentation* = 15 columns is repeatedly the best choice by fragmentation, rather than, say, *fragmentation* = 9 or 18 columns. This is in accordance with the biological model. As mentioned earlier, the binding sites in prokaryotes are typically 20 base-pairs long (Stormo and Hartzell, 1989). Some of these positions are less conserved than others. At the low width value, say, *fragmentation* = 9, some positions that are critical to the function of the binding sites and should be counted on when looking for site instances are excluded from the process, thus leading to more random hits. As the width increases, say, *fragmentation* = 15, the situation ameliorates, and the results improve. But at width 18, when some non-critical positions are included in the information content calculation process, this increases the chance that some random background sites, which lack critical positions and should not be considered as a true site, yet since they score above the cutoff value, are included in the final output.

The Gibbs sampler is one of the most popular and accurate packages in the identification of biosequence motifs. The authors maintain a website,

http://bayesweb.wadsworth.org/gibbs/gibbs.html

where the latest Gibbs sampler (GS new) is accessible. We have submitted to it the Stormo CRP sequence dataset over 1,000 times, each time specifying a different parameter setting, in order to find the best performance of the latest Gibbs sampler on this Stormo dataset, as well as its best parameter setting. The best performance and the best parameter setting for the Stormo CRP sequence dataset GS new are shown in Table 2.

It is clear that our combined approach, GS-PC, considerably improves the performance of the 1995 version of Gibbs sampler (GS old). In the case of the Stormo CRP dataset, GS-PC even outperforms the latest Gibbs sampler.

We then try our algorithm on two other datasets, LexA and purR from *E. coli*, using the best parameter setting learned from the Stormo CRP dataset. The results are shown in Table 3. In both datasets, GS-PC performs comparatively to the latest Gibbs sampler.

4. Conclusion and Future Work

We present a novel combined approach in the identification of TFBSs in prokaryotes that effectively reduces the candidate pattern space to be searched, and make feasible an enumerative scheme in this area. To our best knowledge, this is the first such combined approach.

GS-PC considerably improves the performance of the old GS, and is in the cases tested performing comparably to the latest Gibbs sampler, one of the most accurate tools. Compared to alignment-based approaches, enumerative approaches are fast, and therefore GS-PC is especially valuable when handling large datasets.

Table 1: The *cutoff* is set at 0.10, 0.20, …, 0.70, and at each cutoff, the *fragmentation* in the Gibbs site sampler is varied at 9, 12, …, 18. For each combination of a *cutoff* and *fragmentation* value, GS-PC is run 10 times. The best results at each *cutoff* value are presented here.

Cutoff	Fragmentation	Sites (false)	Number of repeats	Expected frequency
0.10	15	22 (3)	7	1.2×10^{-27}
0.20	15	21 (0)	6	9.5×10^{-28}
0.30	15	21 (0)	6	9.5×10^{-28}
0.40	15	19 (0)	6	3.9×10^{-29}
0.50	15	19 (0)	7	3.9×10^{-29}
0.60	15	19 (0)	6	3.9×10^{-29}
0.70	15	17 (1)	4	1.9×10^{-25}

Table 2: The best performance and the corresponding parameter settings of the old and new Gibbs sampler[a], and GS-PC as run the Stormo CRP sequence dataset.

	Parameter setting	Total sites (false)
GS old	Gibbs site sampler, fragmentation = 10	17 (1)
GS new	Gibbs motif sampler, fragmentation = 16, expected number of sites = 10	19 (1)
GS-PC	not applicable	21 (0)

a. Over the following parameter settings: (i) Gibbs Site Sampler or Gibbs Motif Sampler, (ii) fragmentation, non-

fragmentation, number of columns = 10, 12, …, 30, or local search, number of columns (Gibbs new) = 10, 15, …, 30, and (iii) expected number of sites (Gibbs motif sampler or local search) = 10, 20, 30. Each parameter setting was repeated six times.

Table 3: Performance of GS-PC compared to the best performance of the new Gibbs sampler as run on the LexA and purR datasets.

	GS-PC			GS new [a]
	Fragmentation	Cutoff	Total Sites (false)	Total sites (false)
LexA	15	0.30	18 (1)	19 (1)
		0.50	18 (1)	
purR	15	0.30	17 (2)	15 (0)
		0.50	18 (2)	

a. Gibbs motif sampler, fragmentation = 16, expected number of sites = 15.

5. References

[1] T. L. Bailey and C. Elkan (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learn.*, 21, 51-80.

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

[3] I. Jonassen (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, 13, 509–522.

[4] I. Jonassen, J. F. Collins, and D. G. Higgins (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, 4, 1587–1595.

[5] C. E. Lawrence and A. A. Reilly (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins Struct. Funct. Gen.*, 7, 41–51.

[6] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton (1993) Detecting subtle sequence signals a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208–214.

[7] L. A. McCue, W. Thompson, C. S. Carmack, and C. E. Lawrence (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, 12, 1523–1532.

[8] A. F. Neuwald and P. Green (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, 239, 698–712.

[9] A. F. Neuwald, J. Liu, and C. E. Lawrence (1995) Gibbs Motif Sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, 4, 1618–1632.

[10] M. F. Sagot and A. Viari, "Double combinatorial approach to discovering patterns in biological sequences". In: D Hirschberg and G Myers, eds, *Combinatorial Pattern Matching, Lecture Notes in Computer Science (volume 1075)*, pp. 186–208, 1996. Springer-Verlag.

[11] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188, 415-431.

[12] S. Sinha and M. Tompa (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 30, 5549-5560.

[13] H. O. Smith, T. M. Annau, and S. Chandrasegaran, (1990) Finding sequence motifs in groups of functionally related proteins. *Proc. Natl. Acad. Sci. USA*, 87, 826-830.

[14] R. Staden (1989) Methods for discovering novel motifs in nucleic acid sequences. *Comput. Appl. Biosci.*, 5, 293–298.

[15] G. D. Stormo and G. W. Hartzell III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, 86, 1183–1187.

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

[17] J. van Helden, A. Rios, and J. Collado-Vides (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28, 1808-1818.

[18] F. Wolfertstetter, K. Frech, G. Herrmann, and T. Werner, (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithms. *Comput. Appl. Biosci.*, 12, 71–80.