Will You Find Me?

A Critical Evaluation of Motif Finding Programs: BioProspector, MDscan, and Consensus

Wirulda Pootakham

Final Project
Computational Molecular Biology

June 6, 2003

Introduction

Nowadays an increasing number of genomic sequences, including that of human, are publicly available. Computational biology has provided important data-mining tools, allowing us to exploit the sequence information. New algorithms are continually being developed to create programs that will aid in gene prediction, DNA and protein multiple sequence alignment, protein and RNA secondary structure prediction and motif searching. The motif finding programs have become an invaluable tool in studying transcriptional regulation network. These programs can be used to search for the common motifs shared by the upstream regions of genes that are co-regulated and to identify the regulatory signals such as the transcription factor binding sites. A number of motif finding programs have recently been developed, and they are available for public use. The goal of this project is to critically evaluate the performance of three such programs: BioProspector, MDscan, and Consensus. The gold standard that will be used to assess the performance of these three algorithms is the genomic sequence of the bacteriophage T3. The major advantage of using the T3 genome is that the locations of all promoter sequences are mapped. Additionally, the promoter pattern is well-studied, and the consensus sequence has been experimentally determined. It is, however, important to keep in mind that the generalization of these results, under certain circumstances, may not be appropriate. The T3 genome is relatively compact, and about 90% of the genome encodes for proteins. This composition is very different from that of the human genome, for example. Thus, differing results on the performance of these programs may be achieved with input sequences from other organisms.

Bacteriophage T3

Bacteriophage T3 is a relatively small DNA virus that infects *Escherichia coli, Shigella, Salmonella,* and *Pasteurella.* The virion has an icosahedral head and a small tail. The T3 genome is a linear double-stranded DNA of 38,208 base pairs. Most of the sequence encodes for proteins, and phage T3 employs different strategies to maximize the genetic information. These strategies include gene overlap, internal frame-shifts and internal translational re-initiation (Birge, 2000).

The order of the genes on the T3 genome is important for the regulation of virus multiplication. When a virion attaches to a bacterial cell, the DNA is injected in a linear fashion, with the genes on the left end entering first. These genes possess a set of four closely spaced promoters, called class I promoters (*E. coli* promoters A0-A3), that allow them to be transcribed by the host RNA polymerase even before the entire genome enters the cell. The transcribed messenger RNAs are then processed by the host RNaseIII into five smaller mRNA molecules.

One of the early proteins inhibits the host restriction system, and the others stop the action of the host RNA polymerase (Birge, 2000).

An important early protein is a T3 RNA polymerase, which is involved in major transcription processes of the phage. T3 RNA polymerase only uses phage-specific promoters that are distributed along the center portion of the genome. These regions contain class II and class III genes which are involved in DNA metabolism and virion structure and assembly, respectively (Bailey *et al.* 1983). T3 RNA polymerase is a monomeric protein of MW 103,000 that has remarkable template specificity for T3 DNA; even the DNA of closely related bacteriophage T7 is transcribed very poorly (Joho *et al.* 1990). The specificities of T3 RNA polymerases were intensively studied in the 80's, and the consensus sequence of major class III promoters has been experimentally determined (Basu *et al.* 1984; shown below). Because the promoter sequences have been mapped, T3 genome is an ideal model to be used in evaluating the efficiency of the motif finding programs.

5'… $\begin{matrix}T\\A\end{matrix}$ - A - $\begin{matrix}T\\A\end{matrix}$ - T - $\begin{matrix}A\\T\end{matrix}$ - A - C - C - C - T - C - A - C - T- A- A - A - **+1** G - G - G - A **… 3'**

BioProspector

Background

BioProspector (http://bioprospector.stanford.edu/index.html), developed by the Brutlag Bioinformatics Group, is an algorithm for finding sequence motifs from a set of DNA sequences. This program is successful in finding the binding motifs for *Saccharomyces cerevisiae* RAP1, *Bacillus subtilis* RNA polymerase, and *Escherichia coli* CRP. BioProspector is currently under further development so that it can be combined with a microarray clustering program to examine the upstream regions of genes in the same gene expression pattern group and potentially identify the regulatory sequences (Liu *et al.* 2001).

BioProspector adopts the Gibbs sampling approach with the additional improvements in flexibility and sensitivity. Gibbs sampler searches for the most probable motifs and finds the optimal width and number of these motifs in each sequence. In the first step, one sequence from the input is selected to be a left-out sequence, and the rest of the sequences will be used to find an initial guess of the motif. A random start position for the motif is chosen for all sequences except the left-out sequence, and the motif without the left-out sequence is obtained. The goal is to find the most probable pattern shared by all of the sequences by sliding them back and forth until the ratio of the motif probability to the background probability is maximal.

An additional improvement that BioProspector employs is a *threshold sampler*. This adjustment is based on the fact that there may be more than one transcription factor binding site associated with each group of sequences. As a consequence, some input sequences may not have a copy of a particular motif while the other input sequences may have multiple copies. Furthermore, it is plausible that one input sequence contains more than one binding site. Such sequence may require a binding of a homodimer or having two closely spaced binding sites may increase the chance of transcription factor binding.

Method

The annotated T3 genome sequence is available on the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=NC_003298). The promoter sequences are mapped and the list is shown in Table I. Three categories of input sequences will be used to evaluate the performance of the programs. The first category is the annotated T3, T7 and *E. coli* promoter sequences (for details, see below). The second is the intergenic regions in the T3 genome (Appendix D). An intergenic region is defined as a DNA sequence that lies between two annotated coding sequences. The third category of the input is the "genes." Each 'gene' consists of a coding sequence and an intergenic region upstream of that coding sequence (Appendix D). In the case that several coding sequences share one regulatory region, a 'gene' will consist of the intergenic region upstream of the first coding sequence following by the coding sequences.

For each set of input, three types of promoter sequences are included: fourteen T3 promoter sequences, two *E. coli* promoter sequences (early or Class I promoters) and one T7 promoter sequence. Interestingly, one of the miscellaneous feature sequences in the T3 genome is annotated as a T7 promoter. However, it is not used by the T3 RNA polymerase either *in vivo* or *in vitro.* The *E. coli* and T7 promoter sequences are included in the input in order to detect the frequency of false predictions (1 – specificity).

BioProspector takes a file of DNA sequences, in which the motifs are to be found, either in FASTA format or in tab delimited format. It also requests a file containing background sequences, which will be used to determine the background nucleotide distribution. One of the advantages of using the bacteriophage T3 is the small genome size. This allows for the entire genome to be used as background sequences. Since a minimum of ten background sequences is required, the genome sequence is randomly partitioned into 10 shorter sequences of approximately the same size.

Promoter	**Position in the genome**	**Sequence**
E_coli_A0	126-150	agcctaaagtgatgcctaaagtcaa
E_coli_A1	433-472	ttgactttaagttacctttaaggctattat
E_coli_A2	572-601	ttgacaacgcaaggtaacaagtagtaagat
E_coli_A3	683-711	ttgacacatgaagtaagcacggtacgat
T3_phiOL	366-388	tatttaccctcactaaagggaat
T3_phi1.05	5642-5664	cattaaccctcactaacgggaga
T3_phi1.1	5984-6006	agttaaccctcactaacgggaga
T3_phi1.3	6498-6520	taataaccctcactaacaggaga
T3_phi1.5	7683-7705	cattaaccctcactaacaggaga
T3_phi2.5	8834-8856	taattaccctcactaaagggaac
T3_phi3.8	10603-10625	aattaacactcactaaagggaga
T3_phi4.3	12418-12440	aattaaccctcactaacgggaac
T3_phi6.5	17160-17182	aattaaccctcactaaagggaag
T3_phi9	19698-19720	taattaccctcactaaagggaga
T3_phi10	20733-20755	aattaaccctcactaaagggaga
T3_phi11	22395-22417	ctttaaccctcactaacaggagg
T3_phi13	25457-25479	aattaaccctcactaaagggaga
T3_phiOR	37432-37454	cattaaccctcactaaagggaga
T7	32757-32779	taatacgactcactatagggaga

Table I. The promoter sequences in T3 bacteriophage genome. The positions are according to the sequence (accession number NC_003298) from the NCBI genome database.

BioProspector has three motif models: a one-block motif, a two-block motif and a palindrome motif. The nature of the T3 promoter is an ungapped sequence; therefore, only the one-block motif model will be evaluated here. This model requires a user to specify the motif width. The experimentally defined consensus T3 promoter sequence is 21 nucleotide long (Basu *et al.* 1984; see above). Unfortunately, the specified motif length can only be between 5 and 20 nucleotides. I have chosen to test the efficiency of the program on two different motif widths: 10

and 20 nucleotides. The rationale for using a shorter width of 10 nucleotides is to monitor the sensitivity of the program. Because the Gibbs sampling method is stochastic, the program must be run multiple times in order for most, if not all, of the possible alignments to be found. Each run is likely to start with a different initial guess, which will lead to a discovery of different motifs. There will be three trials for each input sequence, and the top three motifs will be reported.

Result

For each motif, a probability matrix is given along with the consensus sequence, which is determined by the most abundant base at each position. The output also indicates the regions (from input sequences) that contribute to the alignment, their starting positions, and their directions (forward or reverse). An example of the output from BioProspector is illustrated in Appendix A. The number of false negatives, false positives, and possible false positives from each motif is summarized in Tables II and III. The classification of each identified sequence (as a false positive or a possible false positive) is performed by comparing the position and the direction of the identified sequence to the expected position/direction of the actual promoter. A false negative is defined as a T3 promoter sequence that is not discovered by the program. Because the number of true positives is known, the sensitivity[1] can be calculated, and it is given in the table instead of the number of false negatives. A false positive is a sequence erroneously identified by the program. The T7 or the *E. coli* promoters are considered true false positives when claimed as T3 promoters. A "possible false positive" is a sequence identified by the program that is neither a known promoter nor a known false positive, but the positions of these sequences are not within 50 base pairs from the start codon, which is the region where promoters are usually found.

As expected, BioProspector correctly identifies T3 promoter sequences and excludes the *E. coli* and T7 promoters when the motif width is set to be 20 (Table II). One of the parameters that BioProspector takes is whether each sequence has at least one copy of the motif. This parameter together with the *threshold sampler* allows for multiple copies of the motif to be identified in one sequence and for sequences without the motif to be excluded. All three trials performed with the 'defined promoter' input accurately identify the promoter regions even though the consensus from one trial is slightly different from consensuses from the other two trials.

When given the intergenic regions as input sequences, BioProspector still performs relatively well. In one of the three trials, it correctly discovers the promoters with a 100% sensitivity although in the other two trials, some of the promoter sequences are not identified. This suggests that the best result can be achieved with the intergenic region input if the program is operated multiple times. It is remarkable that BioProspector does not mistakenly identify the related T7 promoter or other possible false positives. Interestingly, BioProspector does not discover the

[1] Sensitivity (%) = (# true positives * 100)/(# true positives + # false negatives)

exact same motif from two different inputs (i.e. defined promoter input and intergenic region input). The resulting motif from the first input matches to the -17 to +1 while the motif from the intergenic region input reflects the -14 to +4 of the upstream sequence.

The efficiency of the program declines as a higher portion of the input sequences becomes irrelevant. When the genes are given as input, the program finds the correct motif in two trials. In the first trial (Table II), all three motifs converge to give the same consensus sequence, and in this case, 100% of the T3 promoters are accurately identified and T7 and *E. coli* promoters excluded. In the third trial, one of the three motifs is correct while the other two motifs identified are not the part of the promoter pattern. The second trial completely fails to discover the promoter sequences and identifies a number of possible false positives. This result demonstrates that BioProspector has a potential to uncover *bona fide* promoters when run multiple times. However, without *a priori* knowledge of the promoter sequences, it might be difficult to distinguish the false positives from the true ones.

The performance of the program is re-assessed with the narrower motif width of 10 nucleotides in order to determine whether the sub-region of the promoter sequence can be recognized. As expected, BioProspector performs well in identifying the sub-region of the promoters when the defined promoters are given (Table III). A significant drop in performance is observed when the intergenic regions are used as the input, but surprisingly no false positive is identified in any of the three trials. It is somewhat expected to see a striking decrease in the efficiency when the genes are provided as the input. This is most likely due to the increase of the background noise and the smaller motif width. The reduction in width results in an inferior performance of the program, suggesting that the knowledge of the motif width is a prerequisite for a successful search.

Defined promoters:

Trial	Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive[2]
I	Motif 1	2.228	TATTC ACCTT ACACT AAGGT	100	0	0
	Motif 2	2.228	TATTC ACCTT ACACT AAGGT	100	0	0
	Motif 3	2.228	TATTC ACCTT ACACT AAGGT	100	0	0
II	Motif 1	2.885	TCTAC CCTTT AGTGG AGGGT	100	0	0
	Motif 2	2.885	TCTAC CCTTT AGTGG AGGGT	100	0	0
	Motif 3	2.885	TCTAC CCTTT AGTGG AGGGT	100	0	0
III	Motif 1	2.276	TATTC ACCTT ACACT AAGGT	100	0	0
	Motif 2	2.276	TATTC ACCTT ACACT AAGGT	100	0	0
	Motif 3	2.276	TATTC ACCTT ACACT AAGGT	100	0	0

Intergenic regions:

Trial	Motif	Score	Consensus Sequence	Sensitivity (%)	False Positive	Possible false positive
I	Motif 1	2.332	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
	Motif 2	2.332	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
	Motif 3	2.332	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
II	Motif 1	2.491	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
	Motif 2	2.491	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
	Motif 3	2.491	TTAAC CCTCA CTAAA AGGGA	92.85	0	0
III	Motif 1	2.540	TAAAC CCTCA CTAAA AGGGA	100	0	0
	Motif 2	2.540	TAAAC CCTCA CTAAA AGGGA	100	0	0
	Motif 3	2.540	TAAAC CCTCA CTAAA AGGGA	100	0	0

Genes (Intergenic regions + CDS):

Trial	Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
I	Motif 1	2.408	TTAAC CCTCA CTAAA AGGGA	100	0	1
	Motif 2	2.408	TTAAC CCTCA CTAAA AGGGA	100	0	1
	Motif 3	2.408	TTAAC CCTCA CTAAA AGGGA	100	0	1
II	Motif 1	2.572	CCCTT TTAGT GAGGG TTAAA	0	0	16
	Motif 2	2.546	TCTCC CTTTT AGTGA GGGTT	0	0	16
	Motif 3	2.546	TCTCC CTTTT AGTGA GGGTT	0	0	16
III	Motif 1	2.442	TTACC CCTCA CTAAA AGGGA	100	0	2
	Motif 2	2.383	CCCTT TAGTG AGGGG TTAAT	0	0	13
	Motif 3	2.307	TCTCT CCCTT TTAGT GAGGG	0	0	14

Table II: Summary of the outputs from **BioProspector**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes (see text for definitions). A one-block motif model is used with the motif length of 20 nucleotides.

[2] See text for definition.

Defined promoters:

Trial	Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
I	Motif 1	3.158	ACCCT CCACT	100	0	0
	Motif 2	3.133	ACCCT TCACT	100	0	0
	Motif 3	3.048	TAAAG GGTAA	100	2	0
II	Motif 1	3.454	CCCTT TAGTG	100	1	0
	Motif 2	3.043	CACTA ACGGG	100	1	0
	Motif 3	3.043	CACTA ACGGG	100	1	0
III	Motif 1	3.462	CCCTT TAGTG	100	1	0
	Motif 2	3.230	CTTCA CTAAA	100	1	0
	Motif 3	3.169	ACCCT TCACT	92.85	0	0

Intergenic regions:

Trial	Motif	Score	Consensus Sequence	Sensitivity (%)	False Positive	Possible false positive
I	Motif 1	2.860	TTAGT TAGGG	71.42	0	1
	Motif 2	2.786	CCCTC ACTAA	64.28	0	1
	Motif 3	2.785	TTAGT GAGGG	71.42	0	0
II	Motif 1	2.788	TTAGT GAGGG	92.85	0	1
	Motif 2	2.788	TTAGT GAGGG	92.85	0	1
	Motif 3	2.788	TTAGT GAGGG	92.85	0	1
III	Motif 1	2.857	TTAGT GAGGG	92.85	0	1
	Motif 2	2.844	TTAGT GAGGG	85.71	0	1
	Motif 3	2.824	TTAGT GAGGG	85.71	0	0

Genes:

Trial	Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
I	Motif 1	3.062	CCTTT AGTGA	28.57	1	15
	Motif 2	2.941	TGGCT ATGGG	14.28	1	15
	Motif 3	2.934	GGAGA CCACA	14.28	1	13
II	Motif 1	3.009	TCACT AAAGG	21.42	1	15
	Motif 2	2.996	CACTG AGGAC	14.28	2	15
	Motif 3	2.957	TCCCT TTAGT	50	1	7
III	Motif 1	3.217	TCACT AAAGG	50	1	12
	Motif 2	3.209	CCCTT TAGTG	57.14	1	13
	Motif 3	3.176	CCTTT AGTGA	35.75	1	11

Table III: Summary of the outputs from **BioProspector**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes (see text for definitions). A one-block motif model is used with the motif length of 10 nucleotides.

MDscan

Background

Motif Discovery scan or MDscan (http://bioprospector.stanford.edu/MDscan/) was introduced to examine the chromatin immunoprecipitation (ChIP)-array enriched sequences and to search for the DNA motifs representing the protein-DNA interaction sites. Besides combining two widely used motif search approaches, word enumeration and position-specific weight matrix updating, MDscan integrates the ChIP-array ranking information to increase the speed and efficiency. MDscan first uses the word-enumeration method to search for the motifs that are abundant in the top sequences (e.g. highly ChIP-array enriched fragment) to generate candidate motif patterns. It subsequently updates and refines the motifs using the remaining input sequences. Because MDscan enumerates only existing motifs in the top sequences, its search time increases quadratically with respect to the length of the top sequences and linearly with respect to the rest of the sequences. Additionally, MDscan overcomes the inflexible base substitution by using the *m*-match criterion. For example, at least six matches are required for the two 8-mers to be considered "homologous" ($m = 6$ for 8-mers). In other words, two base substitutions are allowed in two homologous 8-mers. The *m* is determined so that the likelihood of two randomly generated oligomers being *m*-matches of each other is less than 0.15% (Liu *et al.* 2002).

It has been shown that MDscan successfully identified the GAL4, RAP1, and MCB motifs. The top motifs discovered by MDscan correspond to the experimentally identified motifs (Liu *et al.* 2002). Apart from motif finding using data from the ChIP-array experiment, MDscan can be used to search DNA motifs in which the subgroup of the sequences contains abundant motif sequences. This would be useful in expanding the list of known promoter sequences. The known promoters can be given to the program as top sequences and previously unidentified sites may be discovered.

Method

Like BioProspector, MDscan was developed by the Brutlag Bioinformatics Group, so the format of the input background sequences is essentially the same (FASTA). The parameters requested by MDscan are the motif width, the number of top sequences in which the motifs are present abundantly, the number of candidate motifs kept for the refinement step, and the number of motifs reported. As for BioProspector, two different motif widths are used to test the programs: 10 and 20 nucleotides, and three top motifs are reported. The number of top sequences is set to be 7 or 14, and the number of candidate motifs is left as a default value of 20. Because MDscan does not use a stochastic method, it is not necessary to run the same input multiple times.

Result

A similar analysis (applied to the results from BioProspector) is performed with each of the motif reported to categorize the sequence as a true positive, a false positive and a possible false

positive (Table IV). An example of the MDscan output is illustrated in Appendix B. When the motif length is 20 nucleotide long, MDscan successfully identifies T3 promoter sequences with a sensitivity approaching 100% when the inputs are either the defined promoters or the intergenic regions. Unlike BioProspector, MDscan does not have the option where users can specify that not every input sequence will contain the motifs. Consequently, a closely related T7 promoter is repeatedly being identified. The unrelated *E. coli* promoters are, however, excluded efficiently as expected.

The number of top sequences does not have any obvious effect on the results when the input sequences are the defined promoters or the intergenic regions. Given the genes as input sequences, MDscan performs better when it uses 14 top sequences. In this case, the sensitivity is 100%, but the number of possible false positives is so high that, in general, it would be very difficult to distinguish the false positives from the true ones. It is interesting to note that the number of possible false positives is larger than the number of the input sequences. This is because more than one motif from each input sequence can be identified.

When the motif width is set to be 10 nucleotides in order to test whether the program can identify the sub-region of the promoter sequences, MDscan only works efficiently with the defined promoter input, regardless of the number of top sequences specified (Table V). It should be noted that when the intergenic regions and genes are provided as inputs, the sensitivity approaches zero, which means that none of the known promoter sequences is discovered. Furthermore, the number of false positives and possible false positives rise dramatically. This indicates that MDscan works more efficiently when the motif width corresponds to the actual length of the motifs to be found, and more false positives are likely to be identified when the motif width is shortened.

MDscan does not perform as well as expected possibly because the input sequences do not have the appropriate characteristic. The top sequences provided to the program should be more abundant in motifs than the rest of the sequences. This is certainly not true with the input provided here. Each of the input sequence has one known T3 promoter motif. Had the input containing top sequences with highly abundant motif be provided, MDscan is expected to outperform BioProspector and Consensus.

Number of Top Sequences = 7

Defined promoters:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.661	TTACC CTTTA CTAAA GGGTA	85.71	1 (T7)	0
Motif 2	2.646	TTTAC CCTTT ACTAA AGGGT	100	1 (T7)	0
Motif 3	2.643	TACCC TTTAC TAAAG GGTAA	100	1 (T7)	0

Intergenic regions:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.052	ATTAA CCCCT CACTA AAGGG	100	1 (T7)	0
Motif 2	2.052	ATTAA CCCCT CACTA AAGGG	100	1 (T7)	0
Motif 3	2.049	ATTAA CCCCT CACTA AAGGG	100	1 (T7)	1

Genes:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	1.970	CCTTA AGGAT AAACC CTAAG	0	11	74
Motif 2	1.947	CTCAC TAAAG GGGAA ACACC	0	8	29
Motif 3	1.940	CCCCT CACTA AAGGG GAAAG	0	1	17

Number of Top Sequences = 14

Defined promoters:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.683	TTACC CTTTA CTAAA GGGTA	100	0	0
Motif 2	2.683	TTACC CTTTA CTAAA GGGTA	100	0	0
Motif 3	2.664	TTTAC CCTTT ACTAA AGGGT	100	0	0

Intergenic regions:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.059	ATTAA ACCCT CACTA AAGGG	100	1	0
Motif 2	2.022	TAAAC CCTCA CTAAA GGGGA	100	0	0
Motif 3	2.021	ATTAA CCCTC ACTAA AAGGG	100	0	0

Genes:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	1.907	CCTTA AGGCT TCTCT TTGAG	100	1	54
Motif 2	1.886	CCCTT AAAGT TAAAC CCTAA	100	13	53
Motif 3	1.879	TCCAT TTGGT TTCCT CTTTA	100	13	44

Table IV: Summary of the outputs from **MDscan**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes. For each input, two different number of top sequences are used: 7 (top) and 14 (bottom). The motif length is 20 nucleotides, and the top three motifs are reported.

Number of Top Sequences = 7

Defined promoters:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	3.026	CCCTT CACTA	100	1 (T7)	0
Motif 2	2.955	AGTGA AGGGT	100	1 (T7)	0
Motif 3	2.979	CACTA AAGGG	100	1 (T7)	0

Intergenic regions:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.558	CCCTA AAGTG	0	0	36
Motif 2	2.496	ACTTA AAGAG	7.14	0	41
Motif 3	2.475	TCACT TAAAG	7.14	0	33

Genes:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.878	AAAGT GAAAA	0	0	68
Motif 2	2.876	GCCTT TAGTG	0	0	69
Motif 3	2.821	AAAGG AGAAA	0	0	57

Number of Top Sequences = 14

Defined promoters:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	3.005	CCCTT CACTA	100	1	0
Motif 2	2.953	AGTGA AGGGT	100	1	0
Motif 3	2.916	CACTA AAGGG	100	1	0

Intergenic regions:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.735	GGGGG GGGGG	0	0	16
Motif 2	2.681	GGGGG GGGGG	0	0	14
Motif 3	2.668	GGGGG GGGGG	0	0	13

Genes:

Motif	Score	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	3.346	ACTCT AAGGG	0	29	123
Motif 2	3.327	ACTCA AAGGG	0	12	108
Motif 3	3.285	ATGGG AGACC	0	14	88

Table V: Summary of the outputs from **MDscan**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes. For each input, two different number of top sequences are used: 7 (top) and 14 (bottom). The motif length is 10 nucleotides, and the top three motifs are reported.

Consensus

Background

Consensus (http://ural.wustl.edu/~jhc1/consensus/html/Html/main.html) is an algorithm for identifying consensus patterns in a set of unaligned DNA sequences. The method is based on a matrix representation of binding site patterns. Each element in the matrix is determined by the frequency of the indicated base occurring at the indicated position. The goal of the method is to find the most significant matrix (the one with the lowest probability of occurring by chance) out of all the matrices formed. The high information content indicates a rarer and a more desirable matrix. The program also estimates the p-value, which is a probability of observing a particular motif in the alignment of random sequences. The expected frequency is then calculated from multiplying the p-values to the number of possible alignments. This allows the comparison of the matrices deriving from differing number of sequences and having different widths (Herzt *et al.* 1990).

The efficiency in identifying the correct motif improves with the number of sequences, and the time required increases only linearly with the number of sequences. The Consensus program has previously been shown to accurately identify the known consensus pattern for the *E. coli* CRP protein (Stormo and Hartzell, 1989).To further demonstrate the robustness of the program, Herzt *et al.* tested it on eleven DNA sequences containing *E. coli LexA* binding sites. The motifs found were consistent with the known consensus sequence, and Consensus could distinguish the generally accepted *LexA* binding sites from other DNA sequences.

Method

Consensus takes a file of sequences in either the FASTA or the Consensus format. If the sequences are given in the FASTA format, it will be converted into a Consensus format internally before program is run. As with BioProspector and MDscan, three categories of input sequences are used to the test Consensus. The parameters requested are the type of sequence, which in this case is DNA, and the width of the motif, which is specified at 10 and 20 nucleotides.

Result

An example of Consensus output is shown in Appendix C. Once again, as for BioProspector and MDscan, the same analysis is carried out with each motif and the result is summarized in Tables VI and VII. Consensus performs extraordinarily well in identifying T3 promoter sequences when the motif width is set at 20. The sensitivity is at 100% for three types of input sequences including the genes. Moreover, the number of the false positives and possible false positives identified is comparable in all three types of inputs, suggesting that the sensitivity that Consensus provides with the gene input is relatively reliable.

Given the defined promoter and intergenic sequence inputs, the efficiency of Consensus is comparable to BioProspector and MDscan. Interestingly, with the gene input, Consensus performs slightly better than BioProspector in terms of sensitivity. It successfully identifies the promoter patterns with all three types of inputs without reporting too many (possible) false positives. Consensus clearly outperforms MDscan in detecting the promoters when the gene sequences are given as input. While MDscan has discovered all of the true positives, it simultaneously identifies about 10 false positives and more than 50 possible false positives, rendering its result insignificant. One of the many motifs discovered by MDscan could be located in the correct promoter sequences simply by chance.

When the motif width is shortened to 10 nucleotides, Consensus still surpasses the performance of BioProspector and MDscan. It identifies all of the true positives with the minimum number of false positives/possible false positives when provided the defined promoter and intergenic region inputs. Though BioProspector does not identify a large number of (possible) false positives, it also does not discover all of the true positives. None of the motifs found by MDscan are true promoter sequences, and in fact, most of them are possible false positives. With the gene input, the performance of Consensus is quite poor. Only a third of the true positives are discovered. Nevertheless, unlike BioProspector and MDscan, Consensus generally does not identify a large number of false or possible false positives. Not surprisingly, among the very few false positives is the T7 promoter because Consensus does not have the option where users can specify that not every input sequence will contain the motifs. As a consequence, a closely related T7 promoter constantly appears as a false positive.

Motif Width: 20

Defined promoters:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.01747E-110	TTAAC CCTCA CTAAA GGGAG	100	1	0
Motif 2	8.3671E-107	TTAAC CCTCA CTAAA GGGAG	100	0	0
Motif 3	3.3313E-105	TTAAC CCTCA CTAAA GGGAG	100	2	0

Intergenic regions:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	4.07317E-98	TTAAC CCTCA CTAAA GGGAG	100	1	0
Motif 2	7.58525E-96	TTAAC CCTCA CTAAA GGGAG	100	0	0
Motif 3	1.9789E-88	TTAAC CCTCA CTAAA GGGAG	100	0	0

Genes:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.70417E-82	ATTAA CCCTC ACTAA AGGGA	100	1	0
Motif 2	9.47717E-82	ATTAA CCCTC ACTAA AGGGA	100	0	0
Motif 3	4.08416E-77	ATTAA CCCTC ACTAA AGGGA	100	1	1

Table VI: Summary of the outputs from **Consensus**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes. The motif width is 20 nucleotides, and the top three motifs are reported.

Motif Width: 10

Defined promoters:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	4.83731E-49	CCCTC ACTAA	100	1	0
Motif 2	8.74494E-49	ACCCT CACTA	100	0	0
Motif 3	8.74494E-49	CCCTC ACTAA	100	0	0

Intergenic regions:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	1.97442E-45	CCCTC ACTAA	100	0	0
Motif 2	1.97442E-45	ACCCT CACTA	100	0	0
Motif 3	8.84878E-45	ACCCT CACTA	100	1	0

Genes:

Motif	E-value	Consensus	Sensitivity (%)	False positive	Possible false positive
Motif 1	2.75637E-33	CCCTC ACTAA	71.42	2	0
Motif 2	2.75637E-33	CCCTC ACTAA	28.57	2	0
Motif 3	3.95187E-32	CCCTC ACTAA	28.57	2	0

Table VII: Summary of the outputs from **Consensus**. Three types of input sequences are evaluated: the defined promoters, the intergenic regions, and the genes. The motif width is 10 nucleotides, and the top three motifs are reported.

Discussion and Conclusion

To critically evaluate the efficiency of three motif finding programs, BioProspector, MDscan and Consensus, the well-studied promoter sequences of the bacteriophage T3 is employed as a gold standard. Three types of the input sequences are tested. The defined promoters are given as "control" input to demonstrate that the programs are capable of identifying T3 promoter patterns. The intergenic regions and genes are more interesting inputs as they would be used in an actual situation. To search for promoter motifs, it would be most appropriate to use the intergenic regions. Nevertheless, genes are used as input sequences to find out whether any program can identify the motifs when the irrelevant sequences (i.e. coding sequences) are introduced. Two different motif widths are used to test the programs: 10 and 20 nucleotides. The T3 promoters are approximately 20 nucleotides long; therefore, the motif width of 20 should be the most optimal in finding the pattern. The width of 10 is also used primarily to assess whether the programs are able to identify sub-regions of the promoter sequence.

Because it is imperative that the actual width (or the best guess) of the promoters be provided to the programs, the discussion will focus mainly on the results obtained when the motif width is 20. As anticipated, all three programs perform well when given the defined promoters as input, achieving a 100% sensitivity. BioProspector and MDscan identify no false positive while Consensus has a few. For the intergenic region input, BioProspector successfully finds the correct motif in one of the three trials and this is without any false positive. MDscan and Consensus also accomplish the 100% sensitivity; however, these two programs identify a few false positives, especially the T7 promoter. A general recommendation is that MDscan and Consensus should be used when every input sequence is likely to contain at least one motif. If the motif is not expected to be found in all the input sequences, BioProspector would be a preferred choice, and it should be run multiple times in order to identify most, if not all, possible motifs.

When the genes are given as input, about 50% of the time, BioProspector would discover all of the true positives and a number of false/possible false positives. The similar result is obtained with MDscan. Surprisingly, Consensus achieves a 100% sensitivity with very few false positives, suggesting that it is the best program to be used when the inputs contain both regulatory regions and the coding regions.

Lastly, when the motif width is 10, all three programs identify the motif in the defined promoter input as expected. It becomes more challenging when the intergenic regions are provided. BioProspector has an average sensitivity of ~80% and has identified very few false positive whereas MDscan shows a very low sensitivity. Consensus appears to be the best in this case, discovering all the true positives and almost no false positives. If the width of the motif is not known and cannot be estimated easily, Consensus would be an appropriate choice to start your search. It is capable of identifying a sub-motif, allowing you to begin with a short motif width, which can be extended subsequently.

It is important to emphasize that the results reported and the conclusions stated here are obtained from the experiment using the bacteriophage T3 genome. These results provide information on the efficiency of the programs and may be used as a guideline for those who would like to use the motif finding programs. About 90% of T3 genome encodes for proteins and only 10% is the intergenic region. This might not be true for other organisms, especially for higher eukaryotes. Thus, these programs may perform differently with input sequences from other organisms. Furthermore, the T3 promoter sequence is highly conserved. If a more degenerate (less conserved) promoter were to be found using these programs, different outcomes can be expected.

The motif finding programs are very useful in identifying the regulatory sequences especially when the genomic sequences are available. This type of program can also be used in combination with the microarray data to examine the regulatory regions upstream from the genes in the same expression group to look for sequence motifs. Furthermore, it can be employed to identify the protein-DNA interaction sites using the data from ChIP-array experiments. The common motifs from the highly ChIP-array-enriched fragments can be discovered. Computational biology will continue to play an important role in providing tools and facilitating the study of protein-DNA interaction as well as transcriptional regulatory network.

Reference

1. Bailey J.N., Klement J.F., and McAllister W.T. Relationship between promoter structure and template specificities exhibited by bacteriophage T3 and T7 RNA polymerases. *Proc. Natl. Acad. Sci. USA* 1983. 80:2814-2818.
2. Basu S., Sarkar P., Adhya S., and Maitra U. Locations and nucleotide sequences of three major class III promoters for bacteriophage T3 RNA polymerase on T3 DNA. *J. Biol. Chem.* 1984. 259:1993-1998.
3. Birge, Edward. Bacterial and Bacteriophage Genetics. Arizona: Springer, 2000.
4. Hertz G.Z., Hartzell G.W., and Stormo G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 1990. 6:81-92.
5. Joho K.E., Gross L.B., McGraw N.J., Raskin C., and McAllister W.T. Identification of a region of the bacteriophage T3 and T7 RNA polymerases that determines promoter specificity. 1990. 215:31-39.
6. Liu X, Brutlag D.L., and Liu J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*. 2001;:127-38.
7. Liu X.S., Brutlag D.L., and Liu J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat Biotechnol*. 2002. 20:835-839.
8. Mount, David. Bioinformatics: Sequence and Genome Analysis. New York: Cold Spring Harbor Laboratory Press, 2001.
9. Stormo G.D. Computer methods for analyzing sequence recognition of nucleic acid. *Annu. Rev. Biophys. Chem.* 1988. 17:241-263.

Appendices

Appendix A: An example of the output from BioProspector. The intergenic regions are used as input sequences. The motif width is 20. The top three motifs are reported.

```
**************************************
*                                    *
*      BioProspector Search Result   *
*                                    *
**************************************

The highest scoring 3 motifs are:

Motif #1:
*****************************
Width (20, 0); Gap [0, 0]; MotifScore 2.332; Segments 13

Blk1    A     C     G     T      Con  rCon Deg  rDeg
1     0.08  0.01  0.08  0.83      T    A    T    A
2     0.38  0.08  0.01  0.53      T    A    W    W
3     0.91  0.01  0.01  0.08      A    T    A    T
4     0.53  0.38  0.01  0.08      A    T    M    K
5     0.01  0.83  0.16  0.01      C    G    C    G
6     0.08  0.83  0.01  0.08      C    G    C    G
7     0.01  0.91  0.01  0.08      C    G    C    G
8     0.01  0.01  0.01  0.98      T    A    T    A
9     0.01  0.98  0.01  0.01      C    G    C    G
10    0.91  0.08  0.01  0.01      A    T    A    T
11    0.08  0.91  0.01  0.01      C    G    C    G
12    0.01  0.31  0.01  0.68      T    A    Y    R
13    0.61  0.01  0.01  0.38      A    T    W    W
14    0.91  0.01  0.01  0.08      A    T    A    T
15    0.76  0.23  0.01  0.01      A    T    A    T
16    0.38  0.23  0.31  0.08      A    T    R    Y
17    0.16  0.01  0.76  0.08      G    C    G    C
18    0.08  0.01  0.83  0.08      G    C    G    C
19    0.08  0.01  0.91  0.01      G    C    G    C
20    0.91  0.01  0.08  0.01      A    T    A    T

> T3 phiOL_Ing/E coli promoters seg 1 f399 TTAACCCTCACTATAAGGGA
> T3 phi1.05_Ing seg 1 f15 TAACCCTTCACTAACGGAGA
> T3 phi1.1_Ing seg 1 f4 TAACGCCTCACTAACGGGAG
> T3 phi1.3_Ing seg 1 f3 GTAACCCTCACCTAACAGGA
> T3 phi1.5_Ing seg 1 f55 TAACCTCTCACTAACAGGGA
> T3 phi2.5_Ing seg 1 f4 TTACGCCTCACTAAATGGGA
> T3 phi4.3_Ing seg 1 f3 TTAACCCTCACCTAACGGGA
> T3 phi6.5_Ing seg 1 f23 TTAACCCTCCACTAAAGGGA
> T3 phi9_Ing seg 1 f4 ATTACCCTCACCTAAAGGGA
> T3 phi10_Ing seg 1 f4 TAACCACTCACTAAAGTGGA
> T3 phi11_Ing seg 1 f69 TAATCCCTCACTAAACAGGA
> T3 phi13_Ing seg 1 f25 TCAACCCTCACTTAAAGGGA
> T3 phiOR_Ing seg 1 f528 TTAACCCTCACTAAAGGTGA
*****************************

Motif #2:
*****************************
```

Width (20, 0); Gap [0, 0]; MotifScore 2.332; Segments 13

Blk1	A	C	G	T	Con	rCon	Deg	rDeg
1	0.08	0.01	0.08	0.83	T	A	T	A
2	0.38	0.08	0.01	0.53	T	A	W	W
3	0.91	0.01	0.01	0.08	A	T	A	T
4	0.53	0.38	0.01	0.08	A	T	M	K
5	0.01	0.83	0.16	0.01	C	G	C	G
6	0.08	0.83	0.01	0.08	C	G	C	G
7	0.01	0.91	0.01	0.08	C	G	C	G
8	0.01	0.01	0.01	0.98	T	A	T	A
9	0.01	0.98	0.01	0.01	C	G	C	G
10	0.91	0.08	0.01	0.01	A	T	A	T
11	0.08	0.91	0.01	0.01	C	G	C	G
12	0.01	0.31	0.01	0.68	T	A	Y	R
13	0.61	0.01	0.01	0.38	A	T	W	W
14	0.91	0.01	0.01	0.08	A	T	A	T
15	0.76	0.23	0.01	0.01	A	T	A	T
16	0.38	0.23	0.31	0.08	A	T	R	Y
17	0.16	0.01	0.76	0.08	G	C	G	C
18	0.08	0.01	0.83	0.08	G	C	G	C
19	0.08	0.01	0.91	0.01	G	C	G	C
20	0.91	0.01	0.08	0.01	A	T	A	T

> T3 phiOL_Ing/E coli promoters seg 1 f399 TTAACCCTCACTATAAGGGA
> T3 phi1.05_Ing seg 1 f15 TAACCCTTCACTAACGGAGA
> T3 phi1.1_Ing seg 1 f4 TAACGCCTCACTAACGGGAG
> T3 phi1.3_Ing seg 1 f3 GTAACCCTCACCTAACAGGA
> T3 phi1.5_Ing seg 1 f55 TAACCTCTCACTAACAGGGA
> T3 phi2.5_Ing seg 1 f4 TTACGCCTCACTAAATGGGA
> T3 phi4.3_Ing seg 1 f3 TTAACCCTCACCTAACGGGA
> T3 phi6.5_Ing seg 1 f23 TTAACCCTCCACTAAAGGGA
> T3 phi9_Ing seg 1 f4 ATTACCCTCACCTAAAGGGA
> T3 phi10_Ing seg 1 f4 TAACCACTCACTAAAGTGGA
> T3 phi11_Ing seg 1 f69 TAATCCCTCACTAAACAGGA
> T3 phi13_Ing seg 1 f25 TCAACCCTCACTTAAAGGGA
> T3 phiOR_Ing seg 1 f528 TTAACCCTCACTAAAGGTGA
****

Motif #3:
****
Width (20, 0); Gap [0, 0]; MotifScore 2.332; Segments 13

Blk1	A	C	G	T	Con	rCon	Deg	rDeg
1	0.08	0.01	0.08	0.83	T	A	T	A
2	0.38	0.08	0.01	0.53	T	A	W	W
3	0.91	0.01	0.01	0.08	A	T	A	T
4	0.53	0.38	0.01	0.08	A	T	M	K
5	0.01	0.83	0.16	0.01	C	G	C	G
6	0.08	0.83	0.01	0.08	C	G	C	G
7	0.01	0.91	0.01	0.08	C	G	C	G
8	0.01	0.01	0.01	0.98	T	A	T	A
9	0.01	0.98	0.01	0.01	C	G	C	G
10	0.91	0.08	0.01	0.01	A	T	A	T
11	0.08	0.91	0.01	0.01	C	G	C	G
12	0.01	0.31	0.01	0.68	T	A	Y	R
13	0.61	0.01	0.01	0.38	A	T	W	W

14	0.91	0.01	0.01	0.08	A	T	A	T
15	0.76	0.23	0.01	0.01	A	T	A	T
16	0.38	0.23	0.31	0.08	A	T	R	Y
17	0.16	0.01	0.76	0.08	G	C	G	C
18	0.08	0.01	0.83	0.08	G	C	G	C
19	0.08	0.01	0.91	0.01	G	C	G	C
20	0.91	0.01	0.08	0.01	A	T	A	T

> T3 phiOL_Ing/E coli promoters seg 1 f399 TTAACCCTCACTATAAGGGA
> T3 phi1.05_Ing seg 1 f15 TAACCCTTCACTAACGGAGA
> T3 phi1.1_Ing seg 1 f4 TAACGCCTCACTAACGGGAG
> T3 phi1.3_Ing seg 1 f3 GTAACCCTCACCTAACAGGA
> T3 phi1.5_Ing seg 1 f55 TAACCTCTCACTAACAGGGA
> T3 phi2.5_Ing seg 1 f4 TTACGCCTCACTAAATGGGA
> T3 phi4.3_Ing seg 1 f3 TTAACCCTCACCTAACGGGA
> T3 phi6.5_Ing seg 1 f23 TTAACCCTCCACTAAAGGGA
> T3 phi9_Ing seg 1 f4 ATTACCCTCACCTAAAGGGA
> T3 phi10_Ing seg 1 f4 TAACCACTCACTAAAGTGGA
> T3 phi11_Ing seg 1 f69 TAATCCCTCACTAAACAGGA
> T3 phi13_Ing seg 1 f25 TCAACCCTCACTTAAAGGGA
> T3 phiOR_Ing seg 1 f528 TTAACCCTCACTAAAGGTGA
****

Appendix B: An example of the output from MDscan. The intergenic regions are used as input sequences. The motif width is 20, and the number of top sequences is 7. The top three motifs are reported.

```
Pm 0.2500 Minimum match (11/20)

Top 3 motifs Wid Score1 Segment Con Deg
Mtf 1 20 2.052 20 ATTAACCCCTCACTAAAGGG AWTWAMCCYYMMCTAAMRGG
Final Motif 1: Wid 20 Score1 2.052 Segment 20
         A     C     G     T     Con  rCon Deg  rDeg
1       69     9     9    13      A    T    A    T
2       33     5     5    57      T    A    W    W
3        5     5     9    81      T    A    T    A
4       53     5     9    33      A    T    W    W
5       77     9     9     5      A    T    A    T
6       29    53     9     9      C    G    M    K
7       21    69     5     5      C    G    C    G
8        5    77    13     5      C    G    C    G
9        5    57     5    33      C    G    Y    R
10       5    29     5    61      T    A    Y    R
11      25    61     5     9      C    G    M    K
12      61    25     5     9      A    T    A    T
13      13    61     9    17      C    G    C    G
14      17     9     5    69      T    A    T    A
15      73     5     5    17      A    T    A    T
16      77     9     5     9      A    T    A    T
17      57    25     9     9      A    T    A    T
18      25    17    49     9      G    C    R    Y
19      13     9    65    13      G    C    G    C
20       5     9    81     5      G    C    G    C
Seq 1 St f397 ATTTGACCCTCACTACAAGG
Seq 2 St f13 ATTAACCCCTCACTAACGGC
Seq 3 St f2 GTTAACACCTCACTAACGGG
Seq 4 St f2 AATTAACCCTCACTTAACAG
Seq 4 St f3 ATTAACCCTCACTTAACAGG
Seq 5 St f53 ATTAACCGCTCACTAACACG
Seq 6 St f2 AATTACACCTCACTAAATGG
Seq 7 St f2 ATTAACAGCTCACTAAAGTG
Seq 8 St f1 AATTAACCCTCACATAACGG
Seq 8 St f2 ATTAACCCTCACATAACGGG
xSeq 9 St f21 AATTAACCCTCTACTAAAGG
xSeq 9 St f22 ATTAACCCTCTACTAAAGGG
Seq 10 St f3 AATTACCCTCACGTAAAGGG
Seq 11 St f2 ATTAACCCCTCACTAAAGTG
Seq 12 St f67 TTTAATCCCTCACTAATCAG
Seq 13 St f23 ATTGAACCCTCACTAAAAGG
Seq 13 St f24 TTGAACCCTCACTAAAAGGG
Seq 14 St f526 CATTAACCCTCACTAAAGGG
Seq 14 St f527 ATTAACCCTCACTAAAGGGG
xSeq 15 St f2 AATACGACTTCACTATAGGG
Mtf 2 20 2.052 20 ATTAACCCCTCACTAAAGGG AWTWAMCCYYMMCTAAMRGG
Final Motif 2: Wid 20 Score1 2.052 Segment 20
         A     C     G     T     Con  rCon Deg  rDeg
1       69     9     9    13      A    T    A    T
2       33     5     5    57      T    A    W    W
3        5     5     9    81      T    A    T    A
```

```
4       53     5     9    33      A    T    W    W
5       77     9     9     5      A    T    A    T
6       29    53     9     9      C    G    M    K
7       21    69     5     5      C    G    C    G
8        5    77    13     5      C    G    C    G
9        5    57     5    33      C    G    Y    R
10       5    29     5    61      T    A    Y    R
11      25    61     5     9      C    G    M    K
12      61    25     5     9      A    T    A    T
13      13    61     9    17      C    G    C    G
14      17     9     5    69      T    A    T    A
15      73     5     5    17      A    T    A    T
16      77     9     5     9      A    T    A    T
17      57    25     9     9      A    T    A    T
18      25    17    49     9      G    C    R    Y
19      13     9    65    13      G    C    G    C
20       5     9    81     5      G    C    G    C
Seq 1 St f397 ATTTGACCCTCACTACAAGG
Seq 2 St f13 ATTAACCCCTCACTAACGGC
Seq 3 St f2 GTTAACACCTCACTAACGGG
Seq 4 St f2 AATTAACCCTCACTTAACAG
Seq 4 St f3 ATTAACCCTCACTTAACAGG
Seq 5 St f53 ATTAACCGCTCACTAACACG
Seq 6 St f2 AATTACACCTCACTAAATGG
Seq 7 St f2 ATTAACAGCTCACTAAAGTG
Seq 8 St f1 AATTAACCCTCACATAACGG
Seq 8 St f2 ATTAACCCTCACATAACGGG
Seq 9 St f21 AATTAACCCTCTACTAAAGG
Seq 9 St f22 ATTAACCCTCTACTAAAGGG
Seq 10 St f3 AATTACCCTCACGTAAAGGG
Seq 11 St f2 ATTAACCCCTCACTAAAGTG
Seq 12 St f67 TTTAATCCCTCACTAATCAG
Seq 13 St f23 ATTGAACCCTCACTAAAAGG
Seq 13 St f24 TTGAACCCTCACTAAAAGGG
Seq 14 St f526 CATTAACCCTCACTAAAGGG
Seq 14 St f527 ATTAACCCTCACTAAAGGGG
Seq 15 St f2 AATACGACTTCACTATAGGG
Mtf 3 20 2.049 20 ATTAACCCCTCACTAAAGGG AWTWAMCCYYMACTAAMRGG
Final Motif 3: Wid 20 Score1 2.049 Segment 20
         A     C     G     T     Con  rCon Deg  rDeg
1       73     9     9     9      A    T    A    T
2       33     5     5    57      T    A    W    W
3        5     5     5    85      T    A    T    A
4       49     5     9    37      A    T    W    W
5       77     9     9     5      A    T    A    T
6       29    49     9    13      C    G    M    K
7       25    65     5     5      C    G    M    K
8        5    73    17     5      C    G    C    G
9        9    57     5    29      C    G    Y    R
10       5    29     5    61      T    A    Y    R
11      25    61     5     9      C    G    M    K
12      65    21     5     9      A    T    A    T
13      13    65     9    13      C    G    C    G
14      13     9     5    73      T    A    T    A
15      69     5     5    21      A    T    A    T
16      77     9     5     9      A    T    A    T
17      53    29     9     9      A    T    M    K
```

```
18      29    17    45     9      G    C    R    Y
19      13     9    65    13      G    C    G    C
20       5     9    81     5      G    C    G    C
Seq 1 St f397 ATTTGACCCTCACTACAAGG
xSeq 1 St b594 ATTTATAGACAACTTACAGG
Seq 2 St f13 ATTAACCCCTCACTAACGGC
Seq 3 St f2 GTTAACACCTCACTAACGGG
Seq 4 St f2 AATTAACCCTCACTTAACAG
Seq 4 St f3 ATTAACCCTCACTTAACAGG
Seq 5 St f53 ATTAACCGCTCACTAACACG
Seq 6 St f2 AATTACACCTCACTAAATGG
Seq 7 St f2 ATTAACAGCTCACTAAAGTG
Seq 8 St f1 AATTAACCCTCACATAACGG
Seq 8 St f2 ATTAACCCTCACATAACGGG
Seq 9 St f21 AATTAACCCTCTACTAAAGG
Seq 9 St f22 ATTAACCCTCTACTAAAGGG
Seq 10 St f3 AATTACCCTCACGTAAAGGG
Seq 11 St f2 ATTAACCCCTCACTAAAGTG
Seq 12 St f67 TTTAATCCCTCACTAATCAG
Seq 13 St f23 ATTGAACCCTCACTAAAAGG
Seq 14 St f526 CATTAACCCTCACTAAAGGG
Seq 14 St f527 ATTAACCCTCACTAAAGGGG
Seq 15 St f2 AATACGACTTCACTATAGGG
Total time 0:0:14.
```

Appendix C: An example of the output from Consensus. The intergenic regions are used as input sequences. The motif width is 20. The top three motifs are reported.

```
COMMAND LINE: ../../Program/consensus/consensus-v6c -L 20 -q 1000 -A a:t c:g
-c0 -pr2 -pt 4 -pf 0 -f
/ural/d/choi/public_html/consensus/user/171.64.70.233/consensus/sequence

***** PID: 9380 *****
L-mer Width: 20
Minimum distance between starting points of words: not relevant
Save the top alignments derived from each intermediate alignment
      Maximum number of matrices to save between cycles: 1000
Status of complementary sequence: IGNORE.
Algorithim options: one match per sequence.
                    Stop only when the maximum number of cycles is reached.
The number of matrices to print.
      Top Matrices saved from each cycle: 4
      Matrices Saved from the last cycle: NONE

***** Sequence information from file
"/ural/d/choi/public_html/consensus/user/171.64.70.233/consensus/sequence".
*****
sequence 1: 1_T3_phiOL_Ing/E_coli_promoters_366(397)
      fragments: 1-900
sequence 2: 2_T3_phi1.05_Ing_12
      fragments: 1-86
sequence 3: 3_T3_phi1.1_Ing_
      fragments: 1-498
sequence 4: 4_T3_phi1.3_Ing_1
      fragments: 1-97
sequence 5: 5_T3_phi1.5_Ing_48
      fragments: 1-78
sequence 6: 6_T3_phi2.5_Ing_1
      fragments: 1-54
sequence 7: 7_T3_phi3.8_Ing_1
      fragments: 1-67
sequence 8: 8_T3_phi4.3_Ing_1
      fragments: 1-46
sequence 9: 9_T3_phi6.5_Ing_20
      fragments: 1-93
sequence 10: 10_T3_phi9_Ing_1
      fragments: 1-104
sequence 11: 11_T3_phi10_Ing_1
      fragments: 1-158
sequence 12: 12_T3_phi11_Ing_61
      fragments: 1-90
sequence 13: 13_T3_phi13_Ing_17
      fragments: 1-84
sequence 14: 14_T3_phiOR_Ing_485(524)
      fragments: 1-647
sequence 15: T7_promoter
      fragments: 1-23

Total number of sequences: 15.
Total number of sequence fragments: 15.
```

```
#**** Information on observed frequency and occurrence of each letter. ****#
#Total number of letters in the input sequences = 3025
A  0.307438; observed occurrence =   930 (letter   1)
C  0.226116; observed occurrence =   684 (letter   2)
G  0.222149; observed occurrence =   672 (letter   3)
T  0.244298; observed occurrence =   739 (letter   4)

PRIOR FREQUENCIES DETERMINED BY OBSERVED FREQUENCIES.
***** Information for the alphabet from the command line. *****
letter   1: A (complement: T)  prior frequency = 0.307438
letter   2: C (complement: G)  prior frequency = 0.226116
letter   3: G (complement: C)  prior frequency = 0.222149
letter   4: T (complement: A)  prior frequency = 0.244298


INFORMATION CONTENT IS CALCULATED USING NATURAL LOGARITHMS (i.e. BASE e).
DIVIDE BY ln(2) = 0.693 TO CONVERT TO BASE 2, WHICH WAS USED IN
PREVIOUS VERSIONS OF THIS PROGRAM.
```

	MATRICES SAVED FOR NEXT CYCLE			
CYCLE	total number	top adjusted information	ln top p-value	ln expected frequency
1	2740	1.7721	0.0000	7.1255
2	676	10.4386	-26.2844	-12.7956
3	850	15.0903	-50.8193	-31.4468
4	755	17.6910	-74.5727	-49.6842
5	817	19.3727	-98.5305	-68.4361
6	733	20.3859	-121.1875	-86.1648
7	767	21.0060	-142.8517	-103.1604
8	742	21.3935	-163.9179	-119.8091
9	712	21.7594	-186.2276	-137.9527
10	774	21.8930	-206.5214	-154.3399
11	709	21.8984	-225.5714	-169.7609
12	752	21.8482	-244.0888	-184.9595
13	668	21.8697	-264.0253	-201.9449
14	757	21.8548	-283.5738	-219.0220
15	875	21.1683	-290.5101	-224.2489

```
INFORMATION CONTENT IS CALCULATED USING NATURAL LOGARITHMS (i.e. BASE e).
DIVIDE BY ln(2) = 0.693 TO CONVERT TO BASE 2, WHICH WAS USED IN
PREVIOUS VERSIONS OF THIS PROGRAM.

THE LIST OF TOP MATRICES FROM EACH CYCLE--sorted by expected frequency (total
of 14):

MATRIX 1
number of sequences = 15
unadjusted information = 23.3358
sample size adjusted information = 21.1683
ln(p-value) = -290.51   p-value = 6.80879E-127
ln(expected frequency) = -224.249   expected frequency = 4.07317E-98
A |   4   0  12  14   0   2   0   0   0  15   0   0  15  14   9   3   0   0  15   4
C |   0   0   0   1  14  13  15   0  15   0  15   0   0   0   6   0   0   0   0   0
G |   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0  12  15  15   0  11
T |  11  15   3   0   0   0   0  15   0   0   0  15   0   1   0   0   0   0   0   0
  1|13  :   1/368   TTTACCCTCACTAAAGGGAA
```

```
  2|5   :   2/14    TTAACCCTCACTAACGGGAG
  3|6   :   3/3     TTAACCCTCACTAACGGGAG
  4|11  :   4/3     ATAACCCTCACTAACAGGAG
  5|9   :   5/50    TTAACCCTCACTAACAGGAG
  6|14  :   6/3     ATTACCCTCACTAAAGGGAA
  7|2   :   7/3     TTAACACTCACTAAAGGGAG
  8|8   :   8/3     TTAACCCTCACTAACGGGAA
  9|7   :   9/22    TTAACCCTCACTAAAGGGAA
 10|12  :  10/3     ATTACCCTCACTAAAGGGAG
 11|15  :  11/3     TTAACCCTCACTAAAGGGAG
 12|10  :  12/63    TTAACCCTCACTAACAGGAG
 13|3   :  13/23    TTAACCCTCACTAAAGGGAG
 14|4   :  14/487   TTAACCCTCACTAAAGGGAG
 15|1   :  15/3     ATACGACTCACTATAGGGAG

MATRIX 2
number of sequences = 14
unadjusted information = 24.1953
sample size adjusted information = 21.8548
ln(p-value) = -283.574   p-value = 7.00605E-124
ln(expected frequency) = -219.022   expected frequency = 7.58525E-96
A |   3   0  11  14   0   1   0   0   0  14   0   0  14  14   8   3   0   0  14   4
C |   0   0   0   0  14  13  14   0  14   0  14   0   0   0   6   0   0   0   0   0
G |   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  11  14  14   0  10
T |  11  14   3   0   0   0   0  14   0   0   0  14   0   0   0   0   0   0   0   0
  1|10  :   1/368   TTTACCCTCACTAAAGGGAA
  2|11  :   2/14    TTAACCCTCACTAACGGGAG
  3|5   :   3/3     TTAACCCTCACTAACGGGAG
  4|14  :   4/3     ATAACCCTCACTAACAGGAG
  5|8   :   5/50    TTAACCCTCACTAACAGGAG
  6|13  :   6/3     ATTACCCTCACTAAAGGGAA
  7|1   :   7/3     TTAACACTCACTAAAGGGAG
  8|7   :   8/3     TTAACCCTCACTAACGGGAA
  9|6   :   9/22    TTAACCCTCACTAAAGGGAA
 10|12  :  10/3     ATTACCCTCACTAAAGGGAG
 11|2   :  11/3     TTAACCCTCACTAAAGGGAG
 12|9   :  12/63    TTAACCCTCACTAACAGGAG
 13|3   :  13/23    TTAACCCTCACTAAAGGGAG
 14|4   :  14/487   TTAACCCTCACTAAAGGGAG

MATRIX 3
number of sequences = 13
unadjusted information = 24.413
sample size adjusted information = 21.8697
ln(p-value) = -264.025   p-value = 2.16414E-115
ln(expected frequency) = -201.945   expected frequency = 1.9789E-88
A |   2   0  11  13   0   1   0   0   0  13   0   0  13  13   7   3   0   0  13   3
C |   0   0   0   0  13  12  13   0  13   0  13   0   0   0   6   0   0   0   0   0
G |   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  10  13  13   0  10
T |  11  13   2   0   0   0   0  13   0   0   0  13   0   0   0   0   0   0   0   0
  1|13  :   1/368   TTTACCCTCACTAAAGGGAA
  2|3   :   2/14    TTAACCCTCACTAACGGGAG
  3|4   :   3/3     TTAACCCTCACTAACGGGAG
  4|1   :   4/3     ATAACCCTCACTAACAGGAG
  5|2   :   5/50    TTAACCCTCACTAACAGGAG
  6|9   :   7/3     TTAACACTCACTAAAGGGAG
  7|10  :   8/3     TTAACCCTCACTAACGGGAA
  8|12  :   9/22    TTAACCCTCACTAAAGGGAA
```

```
 9|11  :  10/3     ATTACCCTCACTAAAGGGAG
10|5   :  11/3     TTAACCCTCACTAAAGGGAG
11|8   :  12/63    TTAACCCTCACTAACAGGAG
12|6   :  13/23    TTAACCCTCACTAAAGGGAG
13|7   :  14/487   TTAACCCTCACTAAAGGGAG
```

Appendix D: The positions of the intergenic and gene input sequences according to the NCBI bacteriophage T3 genomic sequence (NC_003298). The proteins encoded in each of the "gene" are also included. If more than one protein is encoded in that region, only the first one is shown.

Intergenic region	Gene	Protein
1-900	901-1359	S-adenosyl-L-methionine hydrolase
1328-1429	1430-1627	gene 0.6 protein
2883-2975	2976-5630	RNA polymerase
5631-5716	5717-5989	gene 1.05 protein
5984-6081	6082-6222	gene 1.1 protein
6498-6594	6595-7635	DNA ligase
7636-7713	7714-7791	gene 1.5 protein
8834-8887	8888-9586	single-stranded DNA-binding protein
10603-10669	10670-12370	DNA primase/helicase
12418-12463	12466-12678	gene 4.3 protein
17141-17233	17234-17479	gene 6.5 protein
19698-19801	19802-20734	scaffolding protein
20733-20890	20891-22191	minor capsid protein 10B
22335-22424	22425-23015	tail tubular protein A
25437-25520	25521-25931	internal virion protein A
36948-37594	37595-37744	gene 19.5 protein