An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem

Martin Tompa

Department of Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195-2350, U.S.A.
tompa@cs.washington.edu

Abstract

This is an investigation of methods for finding short motifs that only occur in a fraction of the input sequences. Unlike local search techniques that may not reach a global optimum, the method proposed here is guaranteed to produce the motifs with greatest z-scores. This method is illustrated for the Ribosome Binding Site Problem, which is to identify the short mRNA $5'$ untranslated sequence that is recognized by the ribosome during initiation of protein synthesis. Experiments were performed to solve this problem for each of fourteen sequenced prokaryotes, by applying the method to the full complement of genes from each. One of the interesting results of this experimentation is evidence that the recognized sequence of the thermophilic archaea *A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii* may be somewhat different than the well known Shine-Dalgarno sequence.

Keywords: motif, ribosome binding site, Shine-Dalgarno sequence, protein synthesis initiation, archaea.

1 The Ribosome Binding Site Problem

1.1 Motivating Computational Problem

Suppose you are presented with 4000 sequences, each of length 20. You are told that approximately one third of these sequences each contains an instance of an undisclosed pattern of length about 5. To cloud matters further, those 1300 occurrences of some unknown pattern are not identical substrings of length 5, but only approximately equal substrings. Your problem is to identify the pattern.

These particular numbers arise in the problem of identifying the mRNA $5'$ untranslated sequence that is recognized by the ribosome during initiation of protein synthesis in a typical prokaryote. (This will be called the "Ribosome Binding Site Problem".) However, the general problem has obvious applications to finding other types of binding sites in nucleic acid sequences, as well as finding short motifs in protein families. For concreteness, the discussion in this paper will focus on the Ribosome Binding Site Problem.

A number of algorithms to find motifs have been proposed previously. (See, for example, Bailey and Elkan (1995), Fraenkel *et al.* (1995), Galas *et al.* (1985), van Helden *et al.* (1998), Hertz *et al.* (1990), Hertz and Stormo (1995), Hertz and Stormo (1999), Lawrence *et al.* (1993), Rocke and Tompa (1998), and Staden (1989).) Many of these algorithms are designed to find longer and more general motifs than those that arise in certain binding site problems. The price paid for this generality is that many are not guaranteed to find globally optimal solutions, since they employ some form of local search that may end in a local optimum. Other drawbacks for the current purposes are discussed in Section 2.

This work focuses on finding the short, ungapped motifs that are typical of certain binding site problems. The method described is guaranteed to find the most significant such motifs, as measured by their z-scores (see Section 3).

1.2 Initiation of Protein Synthesis in Prokaryotes

In order to understand the Ribosome Binding Site Problem, it is helpful to begin with a short review of the process by which the ribosome binds to mRNA in prokaryotes (Kozak (1983), Lewin (1997)). The ribosome is composed of proteins and rRNA molecules, and is the structure in which mRNA is translated into protein. The mRNA consists of the codons to be translated into amino acids, plus transcribed (from DNA) but untranslated (into protein) regions at both its $5'$ and $3'$ ends.

At the initiation of protein synthesis, the ribosome binds to the mRNA at a region near the $5'$ end of the mRNA called the *ribosome binding site*. This is a region of approximately 30 nucleotides of the mRNA that is protected by the ribosome during initiation. The ribosome binding site is approximately centered on the *translation start site*, which is the beginning of the first codon (usually AUG) that will be translated. That is, the ribosome binding site contains not only the first few codons to be translated, but also part of the $5'$ untranslated region.

The ribosome identifies where to bind to the mRNA at initiation not only by recognizing the first codon, but also by recognizing a short sequence in the $5'$ untranslated region within the ribosome binding site. This short mRNA

Bacillus subtilis	$5'\ldots$ CUGGAUCACCUCCUUUCUA $3'$
Lactobacillus delbrueckii	$5'\ldots$ CUGGAUCACCUCCUUUCUA $3'$
Mycoplasma pneumoniae	$5'\ldots$ GUGGAUCACCUCCUUUCUA $3'$
Mycobacterium bovis	$5'\ldots$ CUGGAUCACCUCCUUUCU $3'$
Aquifex aeolicus	$5'\ldots$ CUGGAUCACCUCCUUUA $3'$
Synechocystis sp.	$5'\ldots$ CUGGAUCACCUCCUUU $3'$
Escherichia coli	$5'\ldots$ UUGGAUCACCUCCUUA $3'$
Haemophilus influenzae	$5'\ldots$ UUGGAUCACCUCCUUA $3'$
Helicobacter pylori	$5'\ldots$ UUGGAUCACCUCCU $3'$
Archaeoglobus fulgidus	$5'\ldots$ CUGGAUCACCUCCU $3'$
Methanobacterium thermoautotrophicum	$5'\ldots$ CUGGAUCACCUCCU $3'$
Pyrococcus horikoshii	$5'\ldots$ CUCGAUCACCUCCU $3'$
Methanococcus jannaschii	$5'\ldots$ CUGGAUCACCUCC $3'$
Mycoplasma genitalium	$5'\ldots$ GUGGAUCACCUC $3'$

Table 1: $3'$ end of 16S rRNA for various prokaryotes

sequence will be called the *SD site*, for reasons to be made clear below. The mechanism by which the ribosome recognizes the SD site is relatively simple base-pairing: the SD site is complementary to a short sequence near the $3'$ end of the ribosome's 16S rRNA.

The SD site was first postulated by Shine and Dalgarno (1974) for *E. coli*. Subsequent experiments demonstrated that the SD site in *E. coli* mRNA usually matches at least 4 or 5 consecutive bases in the sequence AAGGAGG, and is usually separated from the translation start site by approximately 7 nucleotides, although this distance is variable. Numerous other researchers such as Vellanoweth and Rabinowitz (1992) and Mikkonen *et al.* (1994) describe very similar SD sites in the mRNA of other prokaryotes. It is not too surprising that SD sites should be so similar in various prokaryotes, since the $3'$ end of the 16S rRNA of all these prokaryotes is very similar (Mikkonen *et al.* (1994)). Table 1 shows a number of these rRNA sequences. Note their similarity, and in particular the omnipresence of the sequence CCUCCU, complementary to the Shine-Dalgarno sequence AGGAGG.

Because of the great similarity among SD sites in several prokaryotes, many authors use the term "Shine-Dalgarno sequence" to refer to the particular sequence AAGGAGG, or a large subsequence of it. The term "SD site" will be used more generally to mean the short $5'$ untranslated mRNA motif recognized by the particular organism's ribosome. For most bacteria this will in fact be a Shine-Dalgarno sequence (that is, very similar to AAGGAGG). However, the SD site need not look like this for all prokaryotes, despite the fact that the $3'$ ends of their 16S rRNA sequences are so similar. In fact, one of the interesting results of this work is evidence that the SD sites of the thermophilic archaea *A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii* may be somewhat different.

1.3 Correspondence to the Computational Problem

The Ribosome Binding Site Problem, then, is to identify an organism's SD site, given the collection of sequences upstream from its putative genes. How does this correspond to the computational problem described in Section 1.1?

The prokaryotic genomes currently sequenced each contain between a few hundred and several thousand annotated genes, so it would be within the normal range for a newly sequenced genome to have $N \approx 4000$ open reading frames identified as candidate genes. To identify this genome's SD site, one must search for instances of a motif of length about 5 within the 20-mer just $5'$ to the translation start site of each of these N open reading frames. (For brevity, the 20-mers from these positions will be called *upstream sequences*.)

There are several reasons why only a fraction of these N upstream sequences will contain an SD site:

1. Some of the open reading frames might not be real genes.
2. Some of the putative translation start sites within the open reading frames might be incorrectly placed.
3. Many of the genes might be parts of operons, which are multi-gene complexes that are transcribed together, possibly not all containing a ribosome binding site.
4. A number of the genes might employ some completely different translation initiation mechanism not involving an SD site. (See, for example, Fargo *et al.* (1998) and Loechel *et al.* (1991).)

The last aspect of the computational problem of Section 1.1 to be justified is the fact that instances of the motif will match only inexactly. This is so because the hybridization of the mRNA's SD site and the ribosome's 16S rRNA need only occur with free energy below some negative threshold. The rules of RNA binding energy (Lewin 1997) govern here so that, for instance, some strong GC pairs can compensate for a mismatched pair of residues, and GU "wobble pairs" release

free energy, though not as much as Watson-Crick pairs. Independent of this explanation, it is an easily observable fact that the collection of SD sites do match only approximately.

1.4 Contributions

In the solutions to the Ribosome Binding Site Problem to be discussed, no use will be made of known 16S rRNA sequences to streamline the search for the (approximately) complementary SD site. The reason for this is to develop more generally applicable algorithms. As mentioned in Section 1.1, the Ribosome Binding Site Problem is only one instance of many sequence analysis problems that involve the identification of short motifs. (For other binding site problems, the set of candidate sequences might be found by expression level array experiments (Chu *et al.* (1998), Roth *et al.* (1998)), or by footprinting, after which methods very similar to those discussed in this paper could be employed.) Few of these other motif problems have an analogue of the 16S sequence to guide the search. Even for the Ribosome Binding Site Problem in newly sequenced genomes, the 16S rRNA sequence is often only predicted by comparison to previous data (Mikkonen *et al.* (1994)), and in particular its exact $3'$ end may not be known.

However, knowledge of some 16S sequences makes the Ribosome Binding Site Problem an excellent starting point, because they can be used to verify the plausibility of the candidate SD sites found. Furthermore, solving the Ribosome Binding Site Problem is important as a step in the validation of true genes, in the identification of the correct translation start sites, and in the identification of operons. In fact, this work originated for exactly those reasons, in connection with TIGR's *M. tuberculosis* annotation project. Steven Salzberg sent the author upstream sequences from approximately 4000 *M. tuberculosis* open reading frames, with a request to look for the SD site of *M. tuberculosis* in this data. The results would be useful to the annotation project in weeding out open reading frames that do not correspond to genes, in moving incorrect translation start sites among true genes, and in identifying operons.

The remainder of this paper describes an investigation of algorithms for the Ribosome Binding Site Problem. Section 2 discusses some of the previously published methods for finding motifs, and why most seem insufficient for problems such as the Ribosome Binding Site Problem. Section 3 describes a novel algorithm based on the construction of certain Markov chains, and Section 4 discusses experimental results from running this algorithm on the full complement of gene sequences of each of fourteen annotated prokaryotic genomes.

2 Previous Methods for Finding Motifs

2.1 Accounting for Absolute Number of Occurrences

If a short sequence s is to be a motif, the first and most obvious characteristic is that there should be an (approximate) occurrence of s in many of the N sequences. This basic idea is at the foundation of Staden's algorithm for finding motifs (Staden 1989). He creates a table containing the number of occurrences of each k-mer s, where an occurrence allows for a small, fixed number of substitutions of residues in s. One of the criteria he applies before declaring s a motif is that the number of such occurrences of s must exceed some threshold.

s	N_s
ATAAA	1139
AATAA	1108
AAATA	1096
ATTAA	1087
AAGAA	1068
AAAGA	1067
TAAAA	1058
AAAAT	1055
GAAAA	1043
AGGAA	1042
AATTA	1036
AGAAA	1027
TAAAT	1024
TAAAG	1017
AAGGA	1014
TTAAA	1008
AAAAA	1004
GAAAT	993
TATAA	984
AAATT	982

Table 2: Twenty most frequently occurring 5-mers in *H. influenzae*'s upstream sequences. N_s is the number of sequences containing s, allowing up to one substitution. The SD site should be a subsequence of TAAGGAGGTGATCCAA.

To understand why this criterion alone is insufficient, consider Table 2, which shows the 20 most frequently occurring 5-mers among *H. influenzae*'s upstream sequences, allowing up to 1 substitution. In this and all subsequent examples, the input is composed of an upstream sequence of length 20 for each annotated gene. In the case of *H. influenzae*, the number of upstream sequences is $N \approx 1700$.

H. influenzae is an AT-rich genome, and this is true even for the N upstream sequences, for which the nucleotide frequencies are 41% A, 12% C, 18% G, 29% T. Table 2, rather than listing likely SD sites, largely reflects the nucleotide biases in the input. The $3'$ end of the 16S rRNA of *H. influenzae* (see Table 1) confirms that, with one exception, these frequent subsequences are not the SD sites sought: the SD site should be a subsequence of TAAGGAGGTGATCCAA.

2.2 Accounting for Background Distribution

The argument in Section 2.1 illustrates the well known fact that the identification of motifs must take the nucleotide background frequencies into account. A popular way that motif-finding algorithms have done this is to use the "in-

formation content" (also called "relative entropy") of the motif. (See, for instance, Fraenkel *et al.* (1995), Hertz *et al.* (1990), Hertz and Stormo (1995), Hertz and Stormo (1999), Lawrence *et al.* (1993), Rocke and Tompa (1998), Schneider *et al.* (1986), and Stormo and Hartzell (1989).) Staden (1989), realizing the necessity of taking background frequency into account, also used relative entropy as a second criterion for selecting motifs.

The relative entropy of a motif is defined as follows. Suppose that the motif has length k, and has approximate occurrences in a subset S of the N input sequences. Then the relative entropy of this motif is defined to be

$$\sum_{j=1}^{k} \sum_{r \in \{A,C,G,T\}} p_{r,j} \log_2 \frac{p_{r,j}}{b_r},$$

where $p_{r,j}$ is the frequency with which residue r occurs in position j among the motif occurrences in S, and b_r is the background frequency of residue r. Relative entropy provides a measure of how well-conserved and how unlikely a motif is with respect to the background distribution. In particular, the more different the distribution $\{p_{r,j}\}$ from the background distribution $\{b_r\}$, the higher the relative entropy of position j.

Relative entropy is a good measure for comparing two motifs that have the same number of occurrences (that is, occur in equinumerous subsets of the N input sequences), but not a good measure if the two motifs occur in a vastly different number of sequences. The reason for this is that relative entropy does not take into account the absolute number of occurrences, depending instead on the relative frequency $p_{r,j}$ of occurrence of each of the nucleotides. For instance, for the uniform background distribution, a perfectly conserved motif that occurs in only a few sequences will have a greater relative entropy than an imperfectly conserved motif that occurs in nearly all the sequences. Because of this, most of the previous applications that use relative entropy depend on the fact that the motif occurs in all, or nearly all, of the N sequences. This is definitely not the case for the Ribosome Binding Site Problem. Realizing this drawback, Staden (1989) and Fraenkel *et al.* (1995) provided separate criteria to be used in conjunction with relative entropy.

2.3 Accounting for Both Criteria

The conclusion to be drawn from this section is that a good measure for comparing motifs such as potential SD sites should take into account both the absolute number of occurrences and the background distribution. One way of doing this might be to multiply the number of occurrences by the relative entropy. Stormo (1990) pointed out that this measure is a log likelihood ratio.

In very recent work, Hertz and Stormo (1999) described a method for estimating the expected frequency of achieving a given relative entropy score. The authors showed how to use this method to compare alignments containing differing numbers of sequences.

Van Helden *et al.* (1998) employed an enumerative method similar in outline to that described in the current paper. Their method, though, does not allow for inexact matches among motif instances, and uses a very different measure of statistical significance of motifs.

3 Statistical Significance of Motif Occurrences

A natural way to take into account both the absolute number of occurrences and the background distribution is to begin as in Section 2.1 by creating a table that, for each k-mer s, records the number N_s of sequences containing an occurrence of s, where an occurrence allows for a small, fixed number c of substitutions of residues in s. A reasonable measure of s as a motif, then, would be based on how unlikely it is to have N_s occurrences if the sequences were drawn at random according to the background distribution.

More specifically, let X be a single random sequence of the specified length L ($L = 20$ for upstream sequences), with residues drawn randomly and independently from the background distribution, or alternatively generated by a Markov chain according to the background dinucleotide distribution. Suppose that p_s is the probability that X contains at least one occurrence of the k-mer s, allowing for c substitutions. Under the reasonable assumption that the N sequences are independent, the expected number containing at least one occurrence of s is Np_s, and its standard deviation is $\sqrt{Np_s(1-p_s)}$. Therefore, the associated z-score is

$$M_s = \frac{N_s - Np_s}{\sqrt{Np_s(1-p_s)}}. \quad (1)$$

M_s is the number of standard deviations by which the observed value N_s exceeds its expectation, and is sometimes called the "z-score", "normal deviate", or "deviation in standard units" (Alder & Roessler 1972). M_s is asymptotically normally distributed, and normalized to have mean 0 and standard deviation 1, making it suitable for comparing different motifs s. Equation (1) will be used as the measure of s as a motif: it measures how unlikely it is to have N_s occurrences of s, given the background distribution, and so incorporates both of the desired criteria.

3.1 The Probability of Occurrence in a Single Sequence

What remains, then, is to determine p_s, the probability that a single random sequence X of length L contains at least one occurrence of s. For the case in which no substitutions are permitted and $|s|$ is not too small, p_s can be approximated well by a certain Poisson process (see Waterman (1995, Section 12.3)). Schbath (1995) extended this to the more general case in which the sequence X is generated by a Markov chain.

Determining p_s *exactly* via generating functions is a well studied problem, even for the case in which substitutions, insertions, and deletions are permitted. Even the simplest case

of unbiased coin flips with no such variations permitted is interesting and somewhat counterintuitive: for the alphabet $\{\mathtt{H}, \mathtt{T}\}$, uniform distribution, and $L = 3$, $p_{\mathtt{HH}} = 3/8$ whereas $p_{\mathtt{HT}} = 1/2$. The cause of the difference is the fact that the pattern $s = \mathtt{HH}$ can overlap itself.

Guibas and Odlyzko (1981, Theorem 3.3) were the first to exhibit generating functions that determine p_s exactly, even in the presence of substitutions, insertions, and deletions, if the characters of X are generated independently. Chrysaphinou and Papastavridis (1990, Theorem 4) extended this result to the case in which X is generated by a Markov chain. The remainder of this paper concentrates on the case of at most one subsitution, no insertions or deletions, and Markov chains of order 1, which are the parameters used in the experimentation described in Section 4.

The algorithm to compute p_s that is implicit in the works of Guibas and Odlyzko (1981) and Chrysaphinou and Papastavridis (1990) requires the computation of a determinant of size $(3|s|+2) \times (3|s|+2)$, most of whose entries are polynomials of degree at least $|s|$. This section concludes with the outline of a more direct and efficient algorithm that was employed in the subsequent experimentation.

Given a pattern string s, construct a deterministic finite automaton M that accepts those strings containing a substring that matches s with at most one substitution, as follows. M has a state for every string u that matches a prefix of s with at most one substitution. The transition from u on input character σ is to that state corresponding to the longest suffix of $u\sigma$, that is, the longest suffix that agrees with some prefix of s allowing at most one substitution. M has $1.5|s|^2 + O(|s|)$ states, and can be constructed in time $O(|s|^2)$ (Gusfield 1997, Theorem 3.4.1).

Given the transition probabilities a_{ij} of the Markov chain G that generates the random sequence X, transform M into a Markov chain M' (not to be confused with G) by assigning transition probability a_{ij} to those transitions of M labeled j out of those states whose corresponding string u ends with the character i. The desired probability p_s is given by the probability, in M', of going from the start state to the accepting state in $|X|$ steps. This can be computed in time $O(|X| \cdot |s|^2)$ by exploiting the sparseness of the transition probability matrix of M'. In particular, although that transition matrix has $\Theta(|s|^4)$ entries, each row has only four nonzero entries, so that each of the $|X|$ matrix-vector product can be computed in time $O(|s|^2)$.

4 Results

The algorithm of Section 3 was applied to the upstream sequences from each of fourteen prokaryotic genomes, ten of which are bacteria and four archaea. Nine of the ten bacterial genomes showed a strong predominance of a standard Shine-Dalgarno sequence consisting of most of AAGGAGG. For example, Tables 3 - 5 show the highest scoring sequences found in *H. influenzae*, *B. subtilis*, and *E. coli*, respectively. The patterns s in these and subsequent tables have been aligned by hand to aid visualization.

s	N_s	Np_s	M_s
TAAGGAG	311	78.14	26.96
AAGGAGA	357	116.1	23.16
CTAAGGA	223	56.69	22.46
ATAAGGA	375	130.4	22.29
AGGAGAA	343	114	22.19
GTAAGGA	225	62.39	20.97
TAAGGAC	210	55.87	20.96
TTAAGGA	356	134.5	19.9
GAGGAAA	336	123.8	19.8
TAAGGAA	387	154.2	19.65
AGGAGTA	202	57.61	19.35
AGGAAAA	475	219.4	18.48
AAGGAGT	223	71.41	18.32
TAAGGAT	261	95.18	17.49
ACAAGGA	243	86.66	17.24
AAAGGAG	313	127.1	17.14
AAGGATA	315	129.2	16.99
GGAGTAA	193	61.86	16.98
AGGAGCA	155	43.96	16.97
AAGGAAC	248	92.73	16.58

Table 3: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *H. influenzae*, whose SD site should be a subsequence of TAAGGAGGTGATCCAA

s	N_s	Np_s	M_s
AAGGAGG	2000	548.7	66.58
AGGAGGT	1408	309.1	65
AAAGGAG	1894	592	57.85
TAAGGAG	1246	331.7	52.37
GGAGGTG	1087	265.6	52.12
AGGAGGC	987	231	51.21
AAGGAGC	988	246.1	48.78
TAGGAGG	1113	313.6	46.98
CAAGGAG	1018	272.3	46.77
CAGGAGG	979	256.1	46.65
AAGGAGT	1140	332.6	46.19
AGGAGGA	1473	510.4	45.54
ATAGGAG	1051	310.6	43.7
AGGAGGG	1456	528.5	43.23
ACAGGAG	916	253.7	42.93
AAAGGCG	823	219.5	41.87
AAAGGTG	1019	310.6	41.81
GAGGTGC	495	99.85	40.04
GAGGTGT	609	142.4	39.79
AACGGAG	802	223.6	39.78

Table 4: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *B. subtilis*, whose SD site should be a subsequence of TAGAAAGGAGGTGATCCAG

s	N_s	Np_s	M_s
TCAGGAG	535	113.2	40.19
TAAGGAG	635	165.5	37.22
CAGGAGT	431	99.7	33.57
AGGAGTA	512	138.3	32.31
AAGGAGT	566	163.6	32.08
ACAGGAG	517	143.9	31.65
CAGGAGA	532	151.1	31.55
ATAAGGA	598	192.6	29.89
CAGGAGG	461	128.4	29.79
CCAGGAG	372	96.54	28.36
TGAGGAG	448	133	27.75
AAGGAGA	658	240.2	27.75
AGGAGAA	650	236.3	27.69
CAGGAGC	356	93.82	27.37
TTAAGGA	543	184.6	26.97
CAAGGAG	486	155.9	26.94
TTCAGGA	417	129.2	25.71
ATCAGGA	419	131.2	25.51
AAGGAGG	541	203.7	24.21
GGAGTAA	433	145.7	24.21

Table 5: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *E. coli*, whose SD site should be a subsequence of TAAGGAGGTGATCCAA

To verify that the scores in such tables are statistically significant, the algorithm was run on simulated input data of the same length and dinucleotide composition. The highest resulting simulated scores corresponding to the *H. influenzae* parameters, for example, were around 4. (See Section 5 for an analytical explanation of this value.)

The exceptional bacterium was *M. genitalium*, almost all of whose highest scoring 7-mers do not complement the $3'$ end of its 16S sequence: see Table 6. Its highest scoring 7-mer only had a score of $M_s = 5.5$; simulated data of the same length and dinucleotide distribution had maximum scores in the range 5 – 7, suggesting that the motifs found in the upstream sequences of *M. genitalium* were of no significance. For comparison, with these same parameters the other thirteen prokaryotic genomes had maximum scores ranging from 12 (*M. pneumoniae*) to 67 (*B. subtilis*). It is interesting that Loechel *et al.* (1991) describe a possible alternative ribosome recognition site specifically in *M. genitalium*.

Table 7 shows the highest scoring sequences found in the related organism *M. pneumoniae*. Note in this case the predominance of the Shine-Dalgarno sequence GGAGG.

Synechocystis sp. was the only other bacterial genome to display *any* non-Shine-Dalgarno motif among its high scoring sequences: the second highest scoring 7-mer was CATCGCC, with a score of $M_s = 16$. Further investigation revealed the nature of this high-scoring sequence. Table 8 shows its highest scoring 7-mers among longer upstream sequences of length 40, allowing *no* substitutions. These strongly reveal the cyanobacterial motif GGC-

s	N_s	Np_s	M_s
CGGTTGT	10	2.078	5.508
CCCGCGC	2	0.1255	5.292
GCTCGGG	4	0.4655	5.183
GCGAGGG	5	0.6967	5.16
TTAATTA	111	71.54	5.069
TAATTAA	119	78.11	5.069
ATCCACG	8	1.629	5.001
CACTGGT	11	2.743	5
ATAATTA	103	65.48	5
GGGGAGG	6	1.027	4.913
CAGGGGT	9	2.046	4.873
CTAATTA	62	34.52	4.86
GGAGATC	10	2.458	4.823
ACCCGCG	3	0.3149	4.787
AGTGATC	15	4.748	4.729
GTAATTA	62	35.07	4.729
GATAACT	33	15.08	4.69
CTAACTG	19	6.948	4.606
ACGGTTG	10	2.598	4.605
TGATCAA	29	12.87	4.559

Table 6: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *M. genitalium*, whose SD site should be a subsequence of GAGGTGATCCAC

s	N_s	Np_s	M_s
GGAGGTG	29	4.338	11.88
GAGGAGG	30	5.153	10.99
CGGAGGT	26	4.521	10.14
ATGGAGG	31	6.431	9.735
AGGAGGT	41	10.25	9.681
GGAGGGA	28	5.524	9.602
GGAGGTC	22	3.816	9.334
AGAGGAG	36	9.376	8.756
CAAGGAG	40	11.17	8.695
GAGAGGA	30	7.367	8.384
AAGGAGG	44	13.75	8.245
GGAGGTA	33	9.006	8.049
AGGAGGA	34	9.523	7.988
GAAGGAG	35	10.18	7.84
AGGAGGG	28	7.27	7.73
AGGAGTT	50	18.02	7.634
GGGGGTA	26	6.763	7.435
AACGGAG	34	10.39	7.381
AAGGAGA	51	19.39	7.283
CAGGAGG	23	5.783	7.19

Table 7: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *M. pneumoniae*, whose SD site should be a subsequence of TAGAAAGGAGGTGATCCAC

s	N_s	Np_s	M_s
CGATCGC	76	1.225	67.56
GCGATCG	63	1.029	61.11
GATCGCC	86	2.699	50.72
GGCGATC	62	2.043	41.96
ATCGCCA	36	3.761	16.63
CGGCGAT	20	1.66	14.24
ATCGCCT	32	3.966	14.09

Table 8: The highest scoring 7-mers, allowing no substitutions, in the length 40 upstream sequences of *Synechocystis sp*.

s	N_s	Np_s	M_s
CCGCACT	76	0.621	95.71
ACCGCAC	72	0.734	83.19
GTGCGGT	46	0.539	61.92
CGCACTT	76	1.63	58.25
AGTGCGG	46	0.845	49.14
AAGTGCG	49	1.62	37.22

Table 9: The highest scoring 7-mers, allowing no substitutions, in the length 40 upstream sequences of *H. influenzae*

GATCGCC known as the Highly Iterated Palindrome, HIP1 (Robinson *et al.* (1995), Karlin *et al.* (1996)).

At these longer upstream sequence lengths and no substitutions, the highest scoring 7-mers of *H. influenzae* also revealed another significant motif, shown in Table 9. These are the Uptake Signal Sequence AAGTGCGGT and its inverted complement ACCGCACTT (Smith *et al.* (1995), Karlin *et al.* (1996)).

Finally we come to the remaining four prokaryotic genomes in the experiment, *A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii*, which are all thermophilic archaea. The highest scoring 7-mers of these four archaea are shown in Tables 10 - 13. What is interesting about these is that their highest scoring sequences display a predominance of the pattern GGTGA or GGTG, which satisfies the requirement of complementarity to a subsequence near the $3'$ end of the 16S rRNA (see Table 1). However, that 16S subsequence is shifted a few nucleotides upstream compared to the bacterial sites discussed above.

Interestingly, Watanabea *et al.* (1997) did a relative entropy analysis of the nucleotide distribution at each small fixed distance from the translation start sites in *M. jannaschii*, and noted corroborating findings: "Although [the relative entropy plot for *M. jannaschii*] is similar to that for bacteria, there are also characteristic features [of *M. jannaschii*]. In the G-rich region corresponding to the SD site, there is a T-rich site. In this region, [A] is lowered ... The G-rich region does not overlap the A-rich region residing in the $5'$ side of the G-rich region." (Watanabea, Gojobori, & Miura 1997, page 16)

s	N_s	Np_s	M_s
GGTGATA	435	60.04	49.28
GGTGACA	243	21.83	47.65
CGGTGAT	180	12.84	46.82
GGTGAGA	324	41.56	44.36
AGGTGAT	382	59.21	42.71
CAGGTGA	208	19.77	42.58
GAGGTGA	289	40.96	39.24
GTGGTGA	244	30.22	39.24
TCGGTGA	154	13.2	38.9
ACGGTGA	168	15.72	38.59
TGGTGAG	250	34.25	37.25
GGTGATC	171	17.55	36.82
AGGTGAC	185	20.37	36.7
TGGTGAT	303	50.36	36.15
GTGATAC	183	21.2	35.37
AGGTGAG	258	39.65	35.09
TGGTGAC	160	17.36	34.41
CTGGTGA	168	18.99	34.39
TAGGTGA	313	58.58	33.84
ATGGTGA	309	58.6	33.3

Table 10: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *M. jannaschii*, whose SD site should be a subsequence of GGAGGTGATCCAG

s	N_s	Np_s	M_s
CGGTGAT	176	27.69	28.4
AGGTGAT	318	78.11	27.73
GGAGGTG	294	76.38	25.43
GGTGATC	211	46.41	24.46
GAGGTGA	291	79.23	24.31
GCGGTGA	129	22.44	22.63
GGTGATT	251	68.96	22.34
CAGGTGA	206	50.61	22.14
AGGTGGT	227	60.71	21.7
GGTGATA	250	71.26	21.59
GGTGATG	231	63.4	21.42
TGGAGGT	259	76.51	21.3
TGGTGAT	229	63.44	21.15
CGGAGGT	157	34.53	21.04
GAGGTGT	207	55.61	20.61
AGGTGCT	147	32.45	20.29
AGGAGGT	279	90.77	20.25
GGGTGAT	257	80.68	20.07
AGGTGTT	195	54.15	19.42
CAGGAGG	233	72.47	19.23

Table 11: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *M. thermoautotrophicum*, whose SD site should be a subsequence of AGGAGGTGATCCAG

s	N_s	Np_s	M_s
GTGGTGA	277	43.91	35.57
GCGGTGA	183	22.66	33.87
GAGGTGA	343	75.53	31.38
CGAGGTG	190	29.2	29.98
GGTGGTG	232	43.7	28.8
GGAGGTG	327	80.69	28
AGGTGAT	282	67.62	26.53
AGGTGGT	264	62.18	26.01
GGTGATA	247	58.03	25.18
GGCGGTG	146	24.4	24.77
GGTGATG	216	49.04	24.14
GAGGTGC	178	36.32	23.73
TGAGGTG	236	59.48	23.24
CGGTGAT	135	23.64	23.05
GAGGTGG	275	78.71	22.58
GGGGTGA	300	93.38	21.91
TGGTGAT	200	49.55	21.65
TGGTGGT	184	44.23	21.25
CGGAGGT	175	41.65	20.88
AGGTGAG	257	77.58	20.78

Table 12: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *P. horikoshii*, whose SD site should be a subsequence of AGGAGGTGATCGAG

s	N_s	Np_s	M_s
GGAGGTG	449	82.2	41.17
CGAGGTG	306	42.47	40.8
GGTGGTG	338	53.11	39.53
GAGGTGG	422	81.63	38.33
GAGGTGA	430	89.39	36.71
AGGTGGT	347	66.87	34.74
AGGTGAT	365	73	34.71
GAGGTGC	301	52.63	34.62
TGAGGTG	376	77.38	34.51
GAGGTGT	335	69.71	32.24
AGAGGTG	381	91.14	30.96
GTGGTGA	294	62.01	29.85
GGCGGTG	208	34.37	29.83
AGGAGGT	377	97.13	28.99
GCGGTGA	210	38.04	28.1
GGTGATA	263	55.72	28.1
CGGAGGT	232	46.44	27.5
GGTGATG	257	60.73	25.51
TGGTGGT	233	54.13	24.59
CGGTGGT	163	29.93	24.48

Table 13: Twenty highest scoring 7-mers, allowing up to one substitution, in the upstream sequences of *A. fulgidus*, whose SD site should be a subsequence of AGGAGGTGATCCAG

5 A Bound on the Maximum Score Among Random Sequences

When the algorithm was run on simulated data of the same length and dinucleotide composition as that of any but the smallest genomes, the maximum 7-mer scores were typically in the range 3.5 – 5. The following probabilistic analysis shows that this is not coincidental.

Theorem 1 *Consider a collection of N random, independent, identically distributed DNA sequences, each one generated by any process whatsoever. Then for any fixed integer k and any $B \geq 4$, when $N \to \infty$ the probability $p(k, B)$ that there exists a sequence s such that $|s| = k$ and $M_s > B$ is less than*

$$\frac{4^k}{2.46e^{B^2/2}}.$$

Proof: The central limit theorem (Birnbaum 1962, Theorem 7.5.5) states that, as $N \to \infty$, $\Pr(M_s \leq B)$ converges uniformly to the cumulative probability function $\Phi(B)$ of the normalized normal variable. Thus,

$$\begin{aligned}
\lim_{N\to\infty} p(k,B) &\leq \lim_{N\to\infty} \sum_s \Pr(M_s > B) \\
&= 4^k(1-\Phi(B)) \\
&= \frac{4^k}{\sqrt{2\pi}} \int_B^\infty e^{-x^2/2} dx \\
&\leq \frac{4^k}{\sqrt{2\pi}} \sum_{x=B}^{\infty} e^{-x^2/2} \\
&< \frac{4^k}{\sqrt{2\pi}} \sum_{i=0}^{\infty} e^{-(B^2/2)-iB} \\
&= \frac{4^k}{\sqrt{2\pi}} \frac{e^{-B^2/2}}{1-e^{-B}} \\
&< \frac{4^k}{2.46e^{B^2/2}},
\end{aligned}$$

the last inequality following from $B \geq 4$. □

For instance, when $k = 7$, the probability that any score exceeds 5 is less than 0.025, and the probability that any score exceeds 6 is less than 1.02×10^{-4}, provided the number N of input sequences is sufficiently large. Contrast this with the fact that, on real genomic data, the maximum score was always at least 11.8, with the exception of *M. genitalium* discussed in Section 4.

6 Conclusion and Further Questions

This paper has presented a method to enumerate the short motifs in its input sequences, together with their exact z-scores, thereby identifying those motifs that are most significant (as measured by z-score). The strengths of the method are that it is exhaustive and exact: all motifs are enumerated,

A	C	G	T	rel. entropy
0.315	0.126	0.194	0.342	0.0262
0.262	0.109	0.382	0.232	0.0852
0.237	0.196	0.475	0.0876	0.28
0.566	0.0995	0.187	0.146	0.182
0.0537	0.0119	**0.903**	0.0308	1.46
0.0657	0.0418	**0.864**	0.0289	1.28
0.0706	0.0139	0.0677	**0.848**	1.07
0.0647	0.0318	**0.827**	0.0766	1.13
0.698	0.0358	0.1	0.163	0.443
0.211	0.0786	0.101	**0.585**	0.351
0.311	0.182	0.168	0.264	0.0149
0.293	0.152	0.19	0.276	0.00527

Table 14: Weight matrix composed from *M. thermoautotrophicum* upstream sequences matching the highest scoring 7-mers, allowing up to one substitution. The bold entries reflect the core GGTGAT.

and their z-scores are computed precisely. Thus, it does not suffer from being heuristic or ending in local optima. A resulting weakness of the method is that the algorithm is not efficient for longer and more complex motifs allowing multiple insertions, deletions, and substitutions.

There are a number of interesting problems and extensions raised by this research:

1. Devise a compelling method to combine the highest scoring sequences so as to produce a single motif, rather than a list of the sequences themselves. It is not difficult to construct a weight matrix from those upstream sequences that match some sequence in an alignment of the highest scoring sequences. For instance, for the *M. thermoautotrophicum* patterns of Table 11, such a weight matrix is shown in Table 14. The positions with relative entropy above 0.3 reflect the core GGTGAT. The problem with this representation is exactly that discussed in Section 2.2: there is no indication that this weight matrix corresponds to 1005 of the 1868 upstream sequences. A matrix derived from fewer but better conserved sequences would have a higher relative entropy and look more impressive, and one derived from more sequences would have a lower relative entropy and look less impressive.
2. Devise an efficient algorithm for accommodating longer patterns with proportionately more substitutions allowed. This requires avoiding the Markov chain construction of Section 3.1 for all but a small fraction of the patterns s that occur approximately in the input sequences.
3. Incorporate into the probability calculation a more accurate RNA binding model than simple substitutions, using the free energy rules (Lewin 1997).
4. Apply the method to other motif problems. Particularly appealing is the problem of finding transcription factor binding sites among genes suggested to be coregulated by expression level array experiments (Chu *et al.* (1998), van Helden *et al.* (1998), Roth *et al.* (1998)).

Acknowledgments

I thank Steven Salzberg for posing the Ribosome Binding Site Problem to me, and for involving me in TIGR's search for SD sites in their *M. tuberculosis* project. Jason Hartline and Dick Hwang provided invaluable assistance both in the analysis and in the experimentation. I thank Shawn Cokus, Phil Green, Anna Karlin, Dick Karp, and Saurabh Sinha for a number of very helpful suggestions. Thanks to an anonymous referee for pointing out the paper by van Helden *et al.* (1998). This material is based upon work supported in part by the National Science Foundation and DARPA under grant DBI-9601046.

References

Alder, H. L., and Roessler, E. B. 1972. *Introduction to Probability and Statistics.* W. H. Freeman and Company, fifth edition.

Bailey, T. L., and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21(1–2):51–80.

Birnbaum, Z. W. 1962. *Introduction to Probability and Mathematical Statistics*. Harper and Brothers.

Chrysaphinou, O., and Papastavridis, S. 1990. The occurrence of sequence patterns in repeated dependent experiments. *Theory of Probability and Its Applications* 35(1):167–173.

Chu, S.; DeRisi, J.; Eisen, M.; Mulholland, J.; Botstein, D.; Brown, P. O.; and Herskowitz, I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.

Fargo, D. C.; Zhang, M.; Gillham, N. W.; and Boynton, J. E. 1998. Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in *Chlamydomonas reinhardtii* chloroplasts or in *Escherichia coli*. *Molecular and General Genetics* 257:271–282.

Fraenkel, Y. M.; Mandel, Y.; Friedberg, D.; and Margalit, H. 1995. Identification of common motifs in unaligned

DNA sequences: application to *Escherichia coli* Lrp regulon. *Computer Applications in the Biosciences* 11(4):379–387.

Galas, D. J.; Eggert, M.; and Waterman, M. S. 1985. Rigorous pattern-recognition methods for DNA sequences: Analysis of promoter sequences from Escherichia coli. *Journal of Molecular Biology* 186(1):117–128.

Guibas, L. J., and Odlyzko, A. M. 1981. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A* 30:183–208.

Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press.

Hertz, G. Z., and Stormo, G. D. 1995. Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In Lim, H. A., and Cantor, C. R., eds., *Proceedings of the Third International Conference on Bioinformatics and Genome Research*, 201–216. World Scientific Publishing Co., Ltd., Singapore.

Hertz, G. Z., and Stormo, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. Forthcoming.

Hertz, G. Z.; Hartzell III, G. W.; and Stormo, G. D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* 6(2):81–92.

Karlin, S.; Mrázek, J.; and Campbell, A. M. 1996. Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Research* 24(21):4263–4272.

Kozak, M. 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiological Reviews* 47(1):1–45.

Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; and Wootton, J. C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–214.

Lewin, B. 1997. *Genes VI*. Oxford University Press.

Loechel, S.; Inamine, J. M.; and Hu, P.-c. 1991. A novel translation initiation region from *Mycoplasma genitalium* that functions in *Escherichia coli*. *Nucleic Acids Research* 19(24):6905–6911.

Mikkonen, M.; Vuoristo, J.; and Alatossava, T. 1994. Ribosome binding site consensus sequence of *Lactobacillus delbrueckii* subsp. *lactis*. *FEMS Microbiology Letters* 116:315–320.

Robinson, N. J.; Robinson, P. J.; Gupta, A.; Bleasby, A. J.; Whitton, B. A.; and Morby, A. P. 1995. Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Research* 23(5):729–735.

Rocke, E., and Tompa, M. 1998. An algorithm for finding novel gapped motifs in DNA sequences. In *RECOMB98: Proceedings of the Second Annual International Conference on Computational Molecular Biology*, 228–233.

Roth, F. P.; Hughes, J. D.; Estep, P. W.; and Church, G. M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* 16:939–945.

Schbath, S. 1995. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics* 1(1):1–16.

Schneider, T. D.; Stormo, G. D.; and Gold, L. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* 188:415–431.

Shine, J., and Dalgarno, L. 1974. The 3′-terminal sequence of *E. coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Science USA* 71:1342–1346.

Smith, H. O.; Tomb, J.-F.; Dougherty, B. A.; Fleischmann, R. D.; and Venter, J. C. 1995. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae Rd* genome. *Science* 269:538–540.

Staden, R. 1989. Methods for discovering novel motifs in nucleic acid sequences. *Computer Applications in the Biosciences* 5(4):293–298.

Stormo, G. D., and Hartzell III, G. W. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Science USA* 86:1183–1187.

Stormo, G. D. 1990. Consensus patterns in DNA. In Doolittle, R. F., ed., *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods in Enzymology*, volume 183. Academic Press. 211–221.

van Helden, J.; André, B.; and Collado-Vides, J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281(5):827–842.

Vellanoweth, R. L., and Rabinowitz, J. C. 1992. The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli in vivo*. *Molecular Microbiology* 6(9):1105–1114.

Watanabea, H.; Gojobori, T.; and Miura, K.-i. 1997. Bacterial features in the genome of *Methanococcus jannaschii* in terms of gene composition and biased base composition in ORFs and their surrounding regions. *Gene* 205:7–18.

Waterman, M. S. 1995. *Introduction to Computational Biology*. Chapman & Hall.