Evaluating Representations for the Shine-Dalgarno Site in *Escherichia coli*

Steven Hampson & Dennis Kibler
Information and Computer Science Department
University of California, Irvine
Irvine, California, U.S.A.

Abstract ***Several methods for identifying individual motif instance by exhaustive evaluation of k-mers ($k \leq 10$) are applied to the pooled Upstream Regions (USR) of all 4289*** *Escherichia coli* ***ORFs. Instances of the Shine-Dalgarno (SD) site are readily identified using these methods. Using these motif instances as starting points, various motif representations and training methods, including several new algorithms, are applied to characterize the complete SD motif. Motif representation languages of increasing power give increasingly better characterizations of the SD motif, permitting more SD sites to be reliably identified. In particular, matrix representation is better than IUPAC which is better than k-mer prototype. However, overly powerful representation also results in suboptimal characterization. A variety of matrix techniques using different representations, objective functions and learning methods yield approximately the same motif, providing evidence for the robustness of the result and the effectiveness of the methods. By these measures, about 1/4 of the ORFs have no better than random SD sites. More biologically realistic motif representation languages might further reduce that fraction.***

Keywords: ribosome binding site, shine-dalgarno, probability & weight matrices

1. Introduction: The Ribosome Binding Site

Protein synthesis is a two-step process. First the DNA is *transcribed* to mRNA by an RNA polymerase complex, and second the mRNA is *translated* to protein by a ribosome, which is a complex of proteins and rRNA. The rates of both steps are controlled by patterns in the Upstream Region of each gene. In order to initiate translation, the ribosome must bind to the mRNA at the start codon (typically AUG) which demarcates the boundary between the translated (coding) region and the transcribed but untranslated region just upstream of it. This area is known as the ribosome binding site (Lewin, 2000). In most bacteria the ribosome recognizes this site based on two sequences in the mRNA: the start codon and a region of about 7 bases approximately 13 bases upstream of it. While highly variable, this 7-base region is approximately complementary to, and is recognized by binding to the 3' tail of the 16S rRNA. It is known as the Shine-Dalgarno (SD) site (Shine, 1974). Translation is possible without an SD site (O'Donnell, 2001; Fargo, 1998) but most genes in *E. coli* have an identifiable SD sequence in the expected location. On the other hand, up to a quarter of *E. coli* genes do not appear to have one.

Based on the 3' tail of the *E. coli* 16S rRNA, the complementary, and presumed optimum SD sequence is TAAGGAG. The central subsequence AGGA is the most highly conserved part (Tompa, 1999; Schultz, 2001; Lukashin, 1998; Besemer, 2001) but the entire 7-mer sequence shows some conservation. The degree of match to that sequence is positively correlated with the rate of translation (Karlin, 2000; Sakai, 2001) and is also correlated to frequency of occurrence, so that over all, the most effective motif instances are also the most frequent.

The SD motif has a number of appealing features as a motif-learning problem: 1) There are approximately 4,289 ORFs in *E. coli*, most of which have an identifiable SD site. Thus, although the "correct" answer is not precisely known, it stands a good chance of being characterized with a high degree of precision; 2) Many bacteria have SD sites, providing a range of related test cases; 3) *E. coli* is well studied so there is a great deal of biological data to complement the statistical data; 4) The most frequent k-mer is complementary to a recognition sequence in the ribosome, which provides a reasonable hypotheses for the optimum SD sequence and the potential for calculating k-mer binding strength; 5) The SD site is highly localized in the USR, providing an additional and independent check on the accuracy of motif detection; 6) Progressively larger USRs can be analyzed, making the problem realistically harder, in order to compare different techniques across a range of problem difficulty; 7) The data is easily available on the Internet; 8) The answer is non-trivial, being highly degenerate and appearing to decay continuously into the statistical background.

The DNA binding sites of transcription-regulating proteins are generally treated as categorical, but this may be in part because the strongest examples have been identified first. Further analysis may eventually show that the continuous-through-background variability of the SD motif is a common phenomena in regulatory binding sites.

This property means that it is difficult to make a simple categorical test as to whether a particular *k*-mer is an SD sequence or not. Instead we favor representations that a) are prototypically centered on the consensus sequence TAAGGAG, b) are localized in the right region of the USR, and c) have a high signal/background ratio, as measured by the ratio of number of SD sites identified in USRs versus those identified in scrambled sequences. It is assumed that for two motifs that match the same number of sites in the real data, the one that matches the fewer random sites is preferable.

2. Motif evaluation

Despite its obvious appeal and popularity, it has been observed that there is no actual evidence that matrix representation is better than more restricted motif representation languages (Pevzner, 2002). Here we address that issue in the context of the SD site and find that matrix representation has clear advantages.

A good motif characterization should have both high coverage (high true positive matches) and high specificity (low false positive matches). Converage and specificity are often inversly related. A more accurate measure might also consider the degree of match, but categorical classification is convenient and more standard. The SD site appears to grade continuously into the random background, so any threshold for binary classification is somewhat arbitrary, but we have chosen a specific comparison point of matching 2000 USRs since there are at least that many relatively well-defined SD sites. Thus, for comparison purposes, coverage is fixed and motifs are evaluated by specificity. Comparison at higher and lower coverage (eg 3000 or 1000 matches) results in lower and higher specificity, but gives qualitatively similar results and is not presented here. In addition, discussing matrix results at 2000 does not preclude using the same motifs to detect a larger or smaller number of instances simply by changing the threshold.

All motif-discovery approaches rely on the fact that the frequency of motif instances in a set of co-regulated USRs is greater than their expected background occurrence rate. This can be explicitly computed as *k*-mer over-representation or indirectly included as a requirement that a motif match a large fraction of the USRs. This approach has proved effective in a wide range of applications, including the SD site. Here we use over-representation to identify individual motif instance, optimize motifs within a representation language, and compare motifs across different representation languages.

Over-representation can be calculated in different ways depending on the choice of background models and the statistical method of computing over-representation (Hampson, 2002). For example, as a background model, a *k*-mer's frequency in the data set might be compared to average *k*-mer frequency in the data set, its frequency in a different but related data set (eg the 20-40 base USR region), or its frequency in a scrambled version of the data. *K*-mer counts in scrambled data can be predicted by a first-order Markov Model (MM1) of the data, so actual randomization and counting is not required. Comparing to a first-order model also suggests further possibilities using higher-order statistics for the background model. Likewise, given a choice of background models, over-representation can be computed in different ways such as ratios, z-scores or binomial probabilities. All of these methods give reasonable and often similar results (Hampson, 2002), so for simplicity we use the ratio of real matches (about 2000 in this case, which can be verified by localization) to matches in a randomized (first-order) version of the data. For a fixed coverage, ratios, z-scores and probabilities are all monotonically related to the number of background matches, so the choice of over-representation function is not important. When applied to probability matrices, the signal/background ratio is correlated with relative entropy, which is a standard matrix comparison metric when a separate background set is not available (Stormo, 2000). Both measure the degree of departure from a first-order model, but the signal/background ratio is also applicable to non-matrix representations and can easily utilize alternative background models.

An even simpler specificity measure is simply to count the number of distinct *k*-mers classified as positive. For a given coverage, the motif with the smallest positive *k*-mer set is probably the most specific. This is less sensitive than measuring the number of matches in an explicit background model, but gives useful information. These counts are also reported when comparing motifs.

An additional issue should be considered in evaluating motif representations: while it may be possible to define a good motif over known instances, it may not perform as well over unknown instances. It is possible for a classifier to *over-fit* the data. This issue is seldom addressed in motif evaluation. Our treatment is deferred until the final section on *k*-mer lists, but the general approach is to divide the data in half, construct a classifier for one half (the training set) and see how well it performs on the other half (the test set). The general result is that all motif representation languages except *k*-mer lists perform equally well on the training and test tests, while *k*-mer lists demonstrate significant over-fitting.

3. Identifying Motif instances

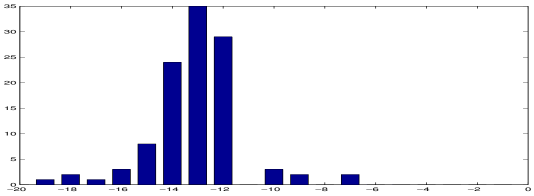
Our methods rely on a few simple statistics. We define M0 as the set of k-mers in the data that exactly match a particular k-mer. M1 is defined as those k-mers in the data that match a given k-mer with exactly one mismatch. M2 is defined similarly. In addition we define C0 as the size of M0, C1 as the size of M1, etc. Only non-overlapping occurrences of a k-mer are counted. C1/3k can be viewed as an nearest-neighbor (M1) estimate of C0.

Given a set of USRs, it is generally assumed that over-represented k-mers of sufficient statistical significance are also of biological significance. Over-representation can be measured in a number of ways but we have focused on two methods that yield similar, subjectively desirable results. One method is to sort k-mers on the ratio C0/C1 where C0 is the number of times the k-mer occurs exactly and C1 is the number of times it occurs with a single mismatch. The other method is to sort k-mers on the probability of observing C0 or more instances based on a Markov model of the data set, which we denote by P(C0 $\geq$ MMn). A scrambled version of the data is frequently used as a comparison set and is equivalent to using a first-order model (MM1). However, we have found that models of order around $k/2$ give results that are most similar to C0/C1. Both methods are empirically effective in identifying over-represented, biologically significant k-mers (Hampson, 2002), but because of their different biases it is often productive to try both.

When applied to the 100-base USRs of 4289 ORFs in *E. coli*, both C0/C1 and P(C0 $\geq$ MM4) identify a number of highly over-represented k-mers, many with distinctive patterns of localization. Many of these appear to be SD instances since they are variations on the predicted optimum 7-mer TAAGGAG and are highly localized in a narrow hill centered about 13 bases upstream. Because of this tight localization, these k-mers can also be identified by sorting on location variance, which has been quite effective in identifying biologically significant k-mers (Hampson, 2002). Sorting on C0 alone identifies some of these strings, but is not particularly selective for them. However, by narrowing the window to a 20-base USR, C0 alone is effective in identifying these k-mers (Table 1), and based on a visual inspection of their localization in the USR (eg Figure 1), at least 98 of the top 100 7-mers fall in this category. Localization could be used as part of the SD site definition, but here it is used only as an independent check on motif definitions based on sequence analysis. Specifically, if a sequence-based motif shows high specificity for the SD region, it is assumed to be detecting real SD sites.

Table 1: Top 20 7-mers sorted on C0 in the 20 base USR. C0/C1 normalized by multiplying by 3k.

	C0	C1	C0/C1	E0	C0/E0
taaggag	110	530	4.36	8	13.75
aaggaga	98	575	3.58	11	8.91
ttaagga	90	456	4.15	8	11.25
aaaagga	82	531	3.23	14	5.86
caggaga	82	457	3.77	6	13.67
aggagaa	81	590	2.88	11	7.36
ataagga	79	536	3.10	10	7.90
aaggagt	76	497	3.21	8	9.50
taaggaa	76	446	3.57	10	7.60
acaggag	74	447	3.48	6	12.33
aaaggag	71	547	2.73	11	6.46
tcaggag	67	470	3.00	4	16.75
aaggaat	66	408	3.40	10	6.60
aacagga	65	409	3.34	8	8.13
aaggaaa	60	529	2.38	14	4.29
aggaata	59	363	3.41	10	5.90
aggataa	59	383	3.24	10	5.90
aggagag	58	389	3.13	8	7.25
acaagga	56	371	3.17	8	7.00
caggagt	56	379	3.10	4	14.00



The average 7-mer would be expected to occur about $(4289 * (1 + 20 - 7))/4^7 = 3.7$ times in this data set, so these k-mers are all highly over-represented. The base frequency in the 20-base USR is not uniform, (ATGC = 33 24 25 18), so it is also useful to compare a k-mer's frequency to its expected frequency in a first-order model of the data. The expected first-order counts and the ratio of observed to expected counts are also shown in Table 1. SD k-mers are highly over-represented by this measure as well, but in a different order. The average C1 value for random k-mers is about 3k times the average C0, or 77, providing additional, independent evidence that these k-mers are exceptional.

It should be noted that even though C1 for these k-mers is highly over-represented, C0 is even more so, producing a high C0/C1 ratio. Likewise for C1 and C2. For the prototypic string TAAGGAG, the average C0 in its M0, M1 and M2 sets is 110, 25, and 9, all in excess of the average value of 3.7. As seen in this and other statistical measures, k-mer frequency is reasonably correlated with similarity to the prototypic SD sequence. This is

Table 2: Context of AGGA: base probability and relative entropy

a	32	47	100	0	0	100	29
t	37	2	0	0	0	0	16
g	16	21	0	100	100	0	45
c	15	30	0	0	0	0	10
RE	0.1	0.2	1.1	1.4	1.4	1.1	0.1

why the C0/C1 ratio selectively identifies prototypic k-mers even for the highly degenerate SD motif. However, this need not be the case since, for example, the best SD sites need not occur at all. Also note that while the best instances (TAAGGAG and its M1 set) are the most frequent, they are present in less than 15% of the USRs. Thus while the presumed best SD sequences are individually the most frequent, the majority of SD sites are distinctly sub-optimum.

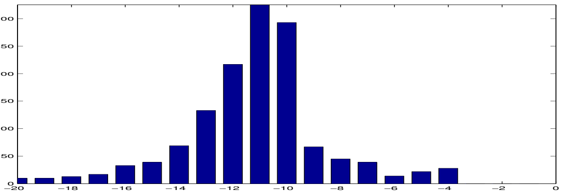


Figure 1: Distribution of 1374 matches for AGGA

The subsequence AGGA is highly conserved and localized (Figure 2), but is not in itself adequate to define the SD site since it occurs only 1374 times in 1340 USRs. AGGA does serve however as a convenient anchor point to measure the size of the SD site. Specifically, its context can be investigated for conserved bases. Base frequency and relative entropy is computed with respect to the first-order distribution for three positions before and after AGGA. The results are reasonably straightforward: only two bases before and one base after the AGGA sequence are noticeably different from the first-order background (Table 2). They demonstrate a considerable degree of variability, but the most frequent base at each position matches the prototypic sequence TAAGGAG, and since they are based on over 1000 examples, should be reasonably accurate. Assuming the choice of AGGA for the central positions does not unduly bias the base frequencies at the other positions, the frequencies at the three context positions should reflect their over-all probabilities in the SD motif. There is only a weak preference for T at the first position and by most measures that position is almost irrelevant. That position pairs with the 3' terminal of the ribosome RNA sequence, so no significant positions before the T would be expected. Note that, as expected, the most frequent base at the second position is A (47%), but C (30%) is actually more informative when compared to their background frequencies (33% and 18%, respectively). This is reflected in Table 1, where TCAGGAG has a higher signal/background ratio (16.75) than TAAGGAG (13.75). An additional G at the end is often stated to be part of the SD site, but does not appear to be favored in this analysis. This is not proof that the SD site for *E. coli* is functionally 7 bases long, but it is convenient and not unreasonable to treat it as such. Analysis at lengths 6 and 8 gave compatible results.

4. Motif Representation

Motif instances are easily identified in this case, but the ultimate goal is generally to give a succinct characterization of the complete motif. Motifs can be represented in various ways, including the following:

1) k-mer prototype 2) IUPAC representation 3) probability matrices 4) weight matrices 5) k-mer list

A k-mer prototype gives a central, prototypic k-mer and the number of mismatches (0-k) that are permitted. If all features are independent and equally important, this can concisely summarize a large set of minor variants on a central theme. The resulting (at least x of k features) prototype is a simple model of the true structure of the motif. There are only $k * 4^k$ possible motifs in this language, so like the set of all 4^k k-mers, it can be evaluated exhaustively for k up to 10 on current work stations. In practice it is only useful to consider M = 0,1,2 settings for $k \leq 10$.

An IUPAC representation permits arbitrary disjunctions at each position, which allows for the fact that some positions may vary in specific ways, while other positions cannot vary at all. All variant features are of equal importance. There are 15^k IUPAC motifs, so exhaustive evaluation is only possible for small k. If desired, IUPAC representation can be combined with the previous prototypic representation by adding a variable mismatch threshold.

Probability matrices are a popular representation language for motifs. If all motif instances are aligned, the frequency of each base can be measured at each position. This 4 x k table is a statistical summary of the instances and is the optimal generating model for those instances, assuming the position probabilities are independent of each other. With the addition of a threshold, a probability matrix can also be used as a detector. The space of matrices is effectively continuous, so exhaustive search is not feasible. Various forms of iterative improvement are generally used to discover local optima. Alternatively, exhaustive k-mer techniques might be used to find most or all motif instances, which are then combined in a single summary table.

A weight matrix, like a probability table, uses a 4 x k table of real numbers. However, rather than summarizing the frequency of each feature, the weight reflects importance for classification purposes. Thus, rather than optimally generating motif instances, the goal is to optimally detect them. An (x of k) prototype can be viewed as a weight matrix of 0s with one 1 in each column and a variable threshold. An IUPAC motif allows any number of 1s and has a threshold of k. Allowing the threshold to vary produces a prototypic IUPAC representation. An arbitrary weight matrix allows all values to be variable and continuous.

A weight matrix can be viewed as a model of binding energy between a sequence and its recognition site, which in turn should be related to the biological effectiveness of the sequence. Each base makes some positive, negative or neutral contribution to binding stability. If the sum of these contributions is greater than some threshold, the binding complex can be considered stable enough to be functional. Binding strength can thus be thought of as the "real" biological motif definition, and optimizing a weight matrix for k-mer classification may approximate that function. In this context, negative weights are not unreasonable, which would obviously not occur in a probability matrix. Likewise, the four weights at a given position in a weight matrix can all be zero or any constant if the position is irrelevant, while probabilities approximate the first-order background and must sum to 1.0. However, a weight matrix is not intrinsically more powerful since any weight matrix can be converted to an equivalent one that conforms to the constraints of a probability matrix (all values positive, columns sum to 1.0). Given an accurate model of binding energy, a optimum weight matrix might in theory be predicted based on the complementary ribosome sequence. However it is not necessarily the case that binding energy can be accurately estimated as the independent contribution of each base in the sequence, and the best current methods for calculating RNA-RNA binding energy utilize neighboring base pairs. In addition, structural issues can influence the relative importance of the various positions.

Feature frequency and feature importance for classification are generally quite close, although even under optimum conditions (independent, uniform feature probability) a probability table over the positive instances of a weight matrix does not necessarily produce the same classification as the weight matrix (Hampson, 1986). In addition, feature frequency can be varied by changing instance frequency without changing the underlying importance. A probability table over positive instances can be used as a classifier or an estimate of binding energy (Stormo, 1998; Benos, 2000) but the representations are not equivalent. Finding the optimum weight matrix suffers from the same search issues as probability matrices.

When used for binary classification, all motif representations define a set of positive k-mers. Restrictions on the representation language restrict the space of possible sets. A k-mer list also defines a positive k-mer set, but makes no assumptions about the underlying structure of the motif and places no restrictions on the motif representation. However, motif strength or biological effectiveness are generally not quantified, so even a comprehensive list is not in itself a complete characterization of the motif. In addition, a list can be quite unwieldy for long, highly degenerate motifs. Too powerful a representation language also increases the dangers of over-fitting the available data.

Table 3: Similarity to TAAGGAG

0	1	2	3	4	5	6	7
1	21	189	945	2835	5103	5103	2187
110	530	1697	4992	12465	19060	15778	5396
8	146	1136	4891	12558	19222	16241	5842
110	25.23	8.98	5.28	4.40	3.74	3.09	2.47
8	6.95	6.01	5.18	4.43	3.77	3.18	2.67
110	525	1414	1657	554	28	1	0

Row 1: Number of mismatches
Row 2: Number of distinct 7-mers
Row 3: Number of matches
Row 4: Number of matches from first-order model
Row 5: Average number of matches per k-mer
Row 6: Average number of matches, first-order model
Row 7: Number of USRs whose best match matches by x

5. K-mer prototype (x of k) representation

TAAGGAG is expected to be the prototypic *E. coli* SD site based on the complementary ribosome sequence, and statistically that appears to be the case. However, there are numerous variations on the theme and the SD site is sometimes simply described as "having some similarity to the sequence TAAGGAG". A simple formalization of this is "matching at least x positions of TAAGGAG". The coverage and specificity of this motif for all values of m is shown in Table 3, which gives the number of distinct k-mers with m mismatches $\binom{(k}{m)*3^m}$, and the total and average number of matches for that set of k-mers. The total and average number of matches in a first-order model is also shown. Last, the number of USRs whose best match with TAAGGAG has exactly m mismatches is shown.

As previously noted, there are 110 exact matches of TAAGGAG, which is greatly in excess of the expected number of matches for a random 7-mer (3.7) or the expected matches based on its first-order base composition (8). It is quite diagnostic of the SD site with a signal/background ratio of 110/8 = 13.7. If one mismatch is allowed, there are C0 + C1 = 640 (potentially overlapping) matches in 635 USRs using 22 different 7-mers, but the ratio drops to 4.3. If two mismatches are allowed, there are 2337 matches in 2049 USRs using 211 different 7-mers with a ratio of 1.8. If three mismatches are allowed, there are 7329 matches in 3706 USRs using 1156 7-mers with a ratio of 1.2. With exactly three mismatches, C3 = 4992 while 4891 would be expected by chance, indicating that there is no better than random selectivity for the SD site in the M3 set. Even if these M3 instances can function as SD sites, they are indistinguishable from the random background.

At the other extreme, over 500 USRs can match no better than three positions of the prototypic sequence, which is less than expected by chance. It is tempting to speculate that ORFs with especially strong or weak SD sites are distinctive in other ways. One speculation was that the presence or absence of SD sites would correlate with operon structure. The expectation was that the initial ORF of an operon would have a strong SD site and than ORFs without identifiable SD sites would necessarily be interior to operons. Somewhat surprisingly, no correlation was found in either case. Apparently the setting of each ORF's SD "rheostat" is completely independent of operon structure.

Clearly, covering a large fraction of the USRs with a *k*-mer prototype leads to an uncomfortably high random hit rate. This is in part because a significant fraction of USRs appear to have no better than random SD sites, but also results from the limitations of the motif representation language. The *k*-mer prototype (at least 5 of TAAGGAG) has good localization of 2107 nonoverlapping matches (Figure 3) in 2049 USRs, using 211 different 7-mers and has a signal/background ratio of 1.8. Using a C at the second position also gives good results: 1998 matches in 1833 USRs with a ratio of 2.1, although the two are not strictly comparable since a smaller number of matches generally results in a higher specificity. Because of the low information content at the first position, a left-shifted version (5 of CAGGAGA) is also reasonable (2132 matches in 1831 USRs, with a ratio of 1.9). An exhaustive search of all (5 of 7) *k*-mer prototypes did not find any with the coverage of TAAGGAG and a better signal/background ratio, but it is not the only reasonable (*x* of *k*) prototype for the SD motif.

6. Probability Matrices

Probability matrices provide a powerful motif representation language, able to capture much of the variability in bases in SD sites. The central problem in producing a probability table is producing an aligned set of motif instances. Various hill-climbing techniques have been used to find a set of *k*-mers in a data set which can be plausibly explained by their resulting probability matrix. However, like hill-climbing in general, they suffer from the problems of multiple local optima and are often rather slow. Because of this, some simply cannot be applied to large data sets.

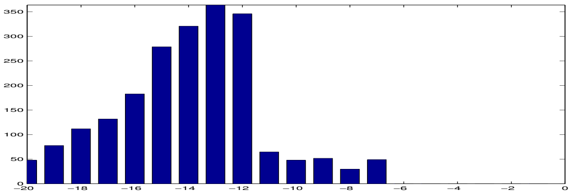


Figure 2: Distribution of 2107 matches for (at least 5 of TAAGGAG)

Table 4: Probability matrix based on M1 neighbors of TAAGGAG

	262	218	148	149	135	146	242
a	27	50	74	11	11	75	31
t	42	3	11	8	4	7	16
g	13	16	6	74	81	14	45
c	18	31	9	7	4	4	8

However, because of the large size of the SD data set, one particularly simple approach for producing a probability table can be used: given a known strong motif instance, assume its M1 neighbors are also motif instances. With this assumption, plus the assumption that the position probabilities are independent, each position can be varied and the frequency of each base measured. Applying this to TAAGGAG provides an estimate of the probability table over all instances (Table 4). For a randomly chosen central *k*-mer, the resulting probability table will be close to the first-order distribution of the data set. The first row gives the total number of patterns over all settings at each position and the numbers below that give the frequency of each base at that position. Thus, for example, with a T in the first position the number of occurrences of the string TAAGGAG is $.42 * 262 = 110$, since there are 110 exact matches. This M1 matrix provides a possible characterization of the SD motif. Similar base strings (e.g. TCAGGAG) yield similar M1 tables.

The M1 table based on the prototypic sequence TAAGGAG is easy to compute, but is only a small-sample estimate of the probability table over all SD instances. More complex techniques can provide more accurate estimates. One standard approach is to iteratively improve a probability table over the complete data set by using an initial table to identify possible motif instances, which then produce a new table, etc (Stormo, 2000). Generalization can be forced by assuming there is a binding site in each USR and defining the probability table over the best match in each USR. This is

Table 5: Iterative probability matrix based on 4058 USR matches

a	36	38	66	0	13	59	34
t	35	5	19	2	6	16	20
g	12	26	0	97	77	16	38
c	17	31	15	1	4	9	8

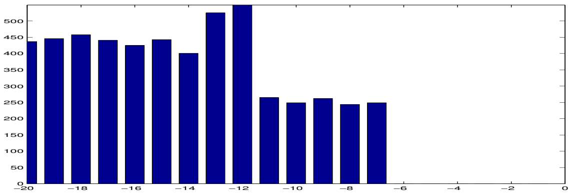


Figure 3: Distribution of 5394 matches for Table 5

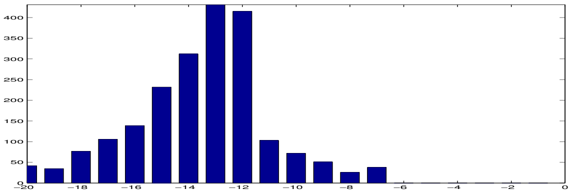


Figure 4: Distribution of 2079 matches for Table 5

often a reasonable assumption, but degrades with the number of USRs without a motif instance. Various *ad hoc* methods have been used to choose the best set of matches. The SD data set has a reasonably high fraction of USRs containing a SD site (at least 3/4) so the issue is not crucial, although including 1/4 random sites is obviously not desirable.

The motif was slid across each USR and each position's match score computed by using the current probability table as a weight vector and computing either an *additive score* by the appropriate dot product, or a *multiplicative score*, corresponding to an estimated probability. Results were similar so only multiplicative ones are reported here. Only best matches that were fully contained in the 20-bp USR were used since those are more apt to be real SD sites. Starting at TAAGGAG and using its M1 probability table as the initial estimate, this process converged on Table 5 in a few seconds. This table is based on 4058 instances. Starting from different initial k-mers similar to TAAGGAG and allowing either 0 or 1 mismatch in the initial probability table gave similar results.

This table is defined over about 4000 instances, but if the threshold is set to match approximately 4000 USRs, localization is very poor (Figure 4) producing 5394 matches in 4058 USRs using 1884 different k-mers with a signal/background ratio of 1.2. Again, it is difficult to cover a large fraction of the USRs without a large number of random matches. If the threshold is set to detect approximately 2000 USRs, localization is much better (Figure 5) producing 2079 matches in 2045 USRs using 126 k-mers and a signal/background ratio of 2.7 - noticeably better than (x of k) representation. Beyond 3000 USRs, localization declines rapidly and the number of matches in the real data and scrambled data are approximately the same, indicating that nearly 1/4 of the USRs do not have SD sites that are distinguishable from a random background.

This result was achieved with a number of starting patterns, but there was also some variability in the final results. One possible factor in the amount of variation in results is that the motif is being trained to fit a significant amount of noise, which by itself produces a large number of different but equally good local optima. Using the best match in all USRs is reasonable, but if a significant fraction of the USRs do not contain identifiable SD sites then only those with the best matches should be used. Consequently, the algorithm was modified so that only the best 2000 USRs were used in defining the probability table on each cycle. A number of slightly different local maxima, with slightly less variability were produced using this method (eg Table 6), in this case with 2025 well-localized matches in 2002 USRs using 113 different 7-mers and with a signal/background ratio of 3.1. Again, beyond 3000, SD sites could not be distinguished from the random background.

A related, but computationally distinct method is to require a certain number of k-mer matches rather than USR matches. For example, assuming that at least 2000 USRs have distinguishable SD sites, all k-mers are ranked by similarity to an initial k-mer or its M1 probability table and the threshold adjusted so the summed C0 values of the positive k-mer set is about 2000. The resulting set of matched k-mers produces a new probability table etc. The computational advantage of this approach is that it is independent of the size of the data set since it can be implemented based entirely on the k-mer set and their C0 counts. For the current analysis, requiring 2000 k-mer matches is very similar to requiring 2000 USR matches, but for motifs with multiple occurrences per USR, the results could be quite different.

A potential disadvantage is that counted motif in-

Table 6: Iterative probability matrix based on best 2002 USR matches

a	37	46	83	0	0	79	32
t	39	0	11	0	0	10	12
g	12	22	0	100	100	11	51
c	12	32	6	0	0	0	5

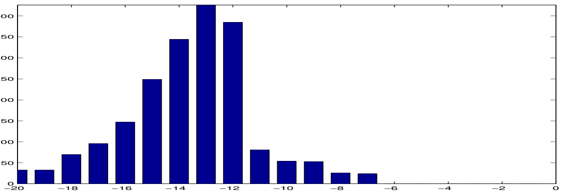


Figure 5: Distribution of 2025 matches for Table 6

Table 7: Iterative probability matrix based on best 2003 *k*-mer matches

a	35	44	79	6	0	88	31
t	36	0	14	0	0	0	13
g	17	22	0	94	100	12	51
c	12	34	7	0	0	0	5

Table 8: Top 60 7-mers for Table 7 and their counts

Top 1-20		Top 21-40		Top 41-60	
taaggag	110	ccaggag	32	taaggac	20
aaaggag	71	tcaggat	15	aaaggac	17
tcaggag	67	acaggat	20	gaaggat	18
acaggag	74	gcaggaa	19	tatggaa	6
taaggaa	76	cgaggag	17	tctggag	45
aaaggaa	54	taaggtg	10	aatggaa	7
tgaggag	34	tgaggat	22	actggag	34
agaggag	16	aaaggtg	15	taaggga	16
tcaggaa	21	agaggat	21	tcagggg	11
acaggaa	46	ggaggaa	39	aaaggga	23
tgaggaa	47	tatggag	16	tacggag	13
agaggaa	29	aatggag	11	acagggg	15
gaaggag	33	taagggg	21	aacggag	24
gcaggag	43	ccaggaa	8	tgaggtg	24
caaggag	48	aaagggg	7	agaggtg	23
taaggat	36	cgaggaa	12	taaagag	17
aaaggat	35	taaggta	15	aaaagag	28
gaaggaa	23	tcaggtg	8	tgtggag	15
ggaggag	23	aaaggta	19	agtggag	12
ccaggag	32	acaggtg	7	tgagggg	8

stances can overlap in the USRs. It is not obvious when or if these should be counted as separate binding sites. This is not a significant issue for the SD motif when starting with TAAGGAG, but it can be for other motifs and other starting instances. Furthermore, there is no bias for maximum USR coverage, which is generally beneficial. Again, in this case it is not a significant issue, but can be under some circumstances.

Starting with TAAGGAG or its M1 table and using multiplicative scoring, this method produces a large number of similar local optima that are also similar to the probability tables obtained by other means. For example, Table 7 has good localization with 2003 matches in 1978 USRs using 119 distinct 7-mers and has a signal/background ratio of 2.8. Similar starting *k*-mers and/or additive scoring gave similar results.

The resultant ranking of *k*-mers by their "motif strength" provides the opportunity for other forms of analysis. For example, the 20 top ranked 7-mers are all highly over-represented and localized, indicating that they are good SD sites. They are distinct, non-overlapping motif instances, unlike sorting on C0, over-representation or location variance. Of the top 100 7-mers, 77 occur more than 10 times and show appropriate localization, indicating that they are real SD sites. The top 60 7-mers are shown in Table 8. Note that while there is some correlation between motif rank and C0, it is far from perfect. The correlation between order and C0 over the top 119 (the positive *k*-mer set) is .67. The degree of correlation reflects how accurately the probability table models the actual instance frequency. High-ranking *k*-mers with low C0 and low ranking *k*-mers with high C0 may point out deficiencies of the motif representation. Correlation is lower using additive similarity, indicating that even though it has similar classification accuracy, it (as expected) is not as accurate in ranking *k*-mers in order of frequency.

7. IUPAC and Weight Matrices

In previous work we extended the calculation of individual *k*-mer over-representation to groups of *k*-mers such as the M1 and M2 shells of a *k*-mer, IUPAC and weight matrix motif representations (Hampson, 2000; Kibler, 2001). As previously observed, over-representation can be calculated in different ways depending on the background model and the statistical method of computing over-representation. Ideally, as a background model a *k*-mer's frequency in the data set could be compared to its frequency in a data set that does not contain the motif in question but is highly similar in all other regards. However, for the SD problem, there is no obvious region that is the same as the 0-20 region but without SD sites, so a first-order background model is employed. Higher-order models were tried but did not appear to be beneficial. Over-representation is computed using z-scores. Thus, unlike probability matrices, the motif is explicitly trained to optimize an objective function much like the signal/background ratio.

As before, non-overlapping matches can be tallied by sliding the motif over the data set, but potentially overlapping matches can be computed more rapidly from the set of *k*-mers and their C0 counts. It is not obvious which is more biologically correct, but for the SD site they gave similar answers. Non-overlapping matches are used here.

Using a first-order comparison set and z-scores to calculate over-representation, a hill-climbing approach to IUPAC optimization is used. Given an initial *k*-mer motif instance, all possible IUPAC codes are tried at each position, over-representation of the resulting positive *k*-mer set calculated, and the code with the biggest value chosen. This is repeated at each position until no further changes are possible. One complete hill-climbing sequence takes several seconds. Randomizing the order that positions are evaluated in often leads to different, but generally similar local optima.

A probability table can be restricted to an IUPAC motif by choosing the most conserved features, but the exact choice of features is not necessarily obvious since it also depends on background feature probability and desired coverage. Optimizing for over-representation takes background frequency into consideration but not desired coverage.

An IUPAC motif can be viewed as a weight matrix with feature weights of 0 or 1 and a threshold of *k*, while an arbitrary weight matrix allows all values to be variable and continuous. However, the goal is the same: to identify a set of *k*-mers with maximum over-representation, starting from a single motif instance. While the instance itself can be used as the starting matrix, better results are generally achieved with a better starting matrix, such as its M1 probability table. This improved results on the average, although the best results of the two initialization methods were about the same. However it is created, the initial weight matrix is incrementally improved by adjusting the weights to include/exclude boundary *k*-mers from the positive set. Boundary *k*-mers are those closest to the hyperplane defined by the current weight matrix and the threshold. Note that while positive/negative *k*-mers are identified based on their degree of over-representation, the weights are adjusted to produce the desired classification of the points, not model their frequency.

Various methods of adjusting the weight matrix to reclassify a boundary point were investigated and the single best method was to try all single-base weight adjustments to include/exclude the given *k*-mer. If including/excluding a point improved the over-representation score of the positive set, the change was accepted and the boundary points recomputed. This was repeated until no further improvements were possible. The 200 closest positive and negative points were considered for adjustment. Increasing the number of boundary patterns considered improves hill-climbing, but takes more time. At some point this time is better spent on multiple restarts. A single hill-climbing episode takes about one minute.

Almost every hill-climbing sequence terminates at a slightly different local maximum. This provides another opportunity to improve the starting matrix. Rather than using the motif instance itself, or its M1 probability table, the start matrix can be set to the sum of previous local maxima. This does not guarantee finding the global optimum, but it does converge on a very high value that is only infrequently found using other initialization techniques. As before, this initialization strategy increased the average local optima score, but not the maximum value found.

Table 9: IUPAC probability table over 2049 matches

a	33	42	100	0	0	62	32
t	35	0	0	0	0	21	20
g	17	29	0	100	100	17	38
c	15	29	0	0	0	0	10

A potential problem with optimizing a motif for over-representation is that the classifier with optimum over-representation need not have the desired coverage. In particular, the classification might be adjusted to include additional, highly over-represented *k*-mers that are almost certainly real SD instances, but might still reduce over-representation of the group as a whole if the additional *k*-mers were significantly less over-represented than the other positive instances. This was frequently observed to be the case, and both IUPAC and weight matrix hill-climbing produced motifs around 1000 positive instances. What is really wanted is the most over-represented *k*-mer set with a given USR coverage, in this case about 2000.

To address this issue, the algorithm was modified in an admittedly *ad hoc* way. In order to encourage the algorithm to find the most over-represented set of about N matches, if the match count was below N, the resulting z-score was multiplied by match/N. This was a minimal modification to bias the evaluation metric towards the desired set size, and worked quite well.

The granularity of IUPAC representation is much coarser than that of a weight matrix so it is harder to closely approximate the desired match count, but the IUPAC motif [ATGC][AGC]AGG[ATG][ATGC] comes close: It has good localization of 2049 matches in 2019 USRs. It uses 144 different 7-mers and has a signal/background ratio of 2.8. The IUPAC motif [ATGC][AGC][ATGC]GGA[ATGC] does almost as well indicating that there are apt to be multiple, equally good representations of the SD motif given almost any representation language.

Weight matrix hill-climbing easily finds weighted combinations of features that are close to the desired coverage, and many similar local optima are produced. Table 10 shows a weight matrix after being converted to probability matrix constraints and the resulting probability table over its positive instances. It has good localization of 2023 matches in 2001 USRs. It uses 106 different 7-mers and has a signal/background ratio of 3.6, which is somewhat better than the best IUPAC or probability matrices. Note that the weights are only roughly correlated with feature frequencies. Nonetheless, when using a multiplicative similarity rule, the resulting probability table can classify all 7-mers the same as the weight matrix, so this motif could in principle be learned by probability matrix techniques. In some cases an additive rule worked while multiplicative did not. There was no clear preference. Despite its higher specificity at 2000 USRs, weight matrices were still not able to distinguish SD instances from the random background beyond 3000.

Other organisms appear to have different cut-offs. For example, in *Synechocystis* 3/4 of the USRs do not have SD sites that could be distinguished from a random background. Requiring a match to only 1/4 of the USRs yields an SD motif at the expected location, but if one requires matching all USRs, the motif is lost (Hayes, 1998).

It is not surprising that a weight matrix which is optimized for over-representation would score well in that regard. Rather it is surprising how similar weight matrix results are to the probability table classifiers in the previous section. The fact that a different representation and objective function yields approximately the same motif is evidence for both the robustness of the result and the effectiveness of the optimization methods. All of these results indicate that SD instances grade into the background at around 3000 USRs, leading to the suspicion that this is close to optimum, at least when characterized with a 4 x 7 matrix.

In addition, the resulting probability table over the positive instances of a weight matrix is usually as effective for instance classification as the weight matrix itself, supporting the direct use of probability tables as classifies. It is possible that weight matrix results could be improved if it was explicitly trained on known positive and negative *k*-mer instances using perceptron training or related algorithms (Hampson, 1999), but such an authoritative classification is not available.

8. Extending matrix representation

Increasing representation power improves specificity and increases the number of SD sites that can be reliably identified. Despite differences in representation and objective functions, the best matrix approaches give approximately the same result, indicating that about 1/4 of the USRs do not have better than random SD sites for *E. coli*. This may well be the case, but it is also possible that more biologically realistic motif representations could do better in detecting SD sites. Three plausible factors were considered as possible extensions to the matrix approach.

If there were more than one type of ribosome, more than one SD matrix might be required, but that does not appear to be the case.

Table 10: Weight matrix and resulting probability table over 2023 matches

a	30	36	42	16	15	46	24
t	30	0	20	16	16	32	24
g	18	28	19	52	53	11	37
c	22	36	19	16	16	11	15

Threshold = -261

a	34	41	78	0	0	83	23
t	37	0	10	0	0	14	16
g	15	24	6	100	100	2	56
c	14	35	6	0	0	1	5

Localization is obviously important for the function of the SD site and might be included as part of the motif strength calculation. For example, it is possible that a greater degree of degeneracy is acceptable at or near the optimum location. This possibility was investigated by creating separate SD probability tables for each starting position in the USR. However, there was no obvious interaction between location and the resulting probability tables, so this option was not pursued.

A third possibility is that there are important inter-feature dependencies that are not captured by the matrix model, which assumes that the base frequencies at each position in the motif are independent of each other. The correlation between motif ranking and motif instance frequency is not perfect at the 2000 USR comparison point (.65), showing that a probability table is a less-than-optimum generator for motif instances and significant high-order dependencies might exist. Initial results from comparing the M1 tables of different motif instances suggested that the probabilities are reasonably independent, but other measures are possible. To look for feature interaction, a 28 x 28 probability table was constructed for the set of motif instances identified at the 2000 USR point. Each column in the table gives the observed frequency of a base at a given motif position for each base setting of all the other positions. If there was no feature interaction for the given motif, all the values in a column would be the same except for sampling error. This generally appears to be the case, with a few exceptions.

One possible interaction is that a G at the second position and a T at the sixth position predict each other since the presence of one more than doubles the probability of the other. The significance of this, if any, is not known.

A general deviation from independence is that choosing a low-probability base at one position usually reduces the chances of choosing a low-probability base at other positions. So, for example, the presence of T or C at the third position, which are low-probability, implies that the sixth position will not be its low-probability choices (T,G), and vice versa. This is not completely unexpected. It is easy to imagine an (x of k) motif that would allow one mismatch but not two, or in keeping with the above results, two mismatches would be possible, but much less frequent than the product of their individual probabilities. The probability table resulting from such an instance set may well correctly classify the instances, but cannot reflect the mutual exclusion of mismatches. Thus, even if the basic assumption of feature independence is violated, a probability table may still be a good instance classifier, even if it is poor at predicting their actual frequency of occurrence.

9. *K*-mer list

The general goal is to optimally characterize the SD site within the constraints of a given motif representation language. The limitations of this approach are apparent if the language is significantly less powerful than the biological recognition mechanism. The advantage is that an appropriate representation language can succinctly represent the actual biological category and has good predictive potential based on a limited number of instances. For optimal results the representation language should mirror the structure of the corresponding biological recognition mechanism as closely as possible. In this regard, the use of increasingly general weight matrices seems justifiable since they have a plausible biological justification. More powerful representation languages permit further improvement in motif specificity but are not necessarily more biologically realistic.

For example, one form of motif representation is to simply list motif instances. This is reasonable if the instances are biologically justified, but is computationally problematic since it places no restrictions at all on the motif representation language, which increases the risk of over-fitting the data. In this context, the SD recognition task simply becomes one of “set coverage” where the goal is to cover as many USRs as possible using as few *k*-mers and first-order matches as possible. This was implemented as a point of comparison. Finding the optimal *k*-mer covering set is NP-complete, but like other optimization problems, reasonable solutions can be constructed by hill-climbing. 7-mers were incrementally chosen to maximize USR coverage while minimizing first-order hits. Specifically, the *k*-mer with the largest value of N/(MM1+3) was always chosen and added to the covering set where N is the number of additional USRs covered and MM1 is the number of first-order matches. Increasing MM1 (by 3 in this case) biases the choice toward *k*-mers with large N, which seemed beneficial.

Using this approach, 2000 USRs could be covered using only 85 *k*-mers with a ratio of 5.13, much better than matrix representation. Localization is good and, not surprisingly, most of the chosen *k*-mers were real SD instances. However, some were not. For example, conserved sequences in homologous ORF families are good for covering certain sets of USRs, but need not coincide with the SD site. Such sequences are generally incompatible with a single matrix representation of real SD instances, and so are appropriately excluded by using a single matrix model. The *k*-mer list motif also shows signs of over-fitting the data. To test for this possibility, the data was divided into two halves. IUPAC and matrix representations trained on one half perform about as well on the other half, but a *k*-mer list does not. Specifically, a *k*-mer list trained to cover 1006 USRs in one half only covered 767 USRs in the other half. The corresponding values for a probability table were 1094 vs 1087, and for a weight matrix 999 vs 962. A *k*-mer list is an extreme example since there are no restrictions at all on the structure of the motif, but other representations such as multi-level neural nets are also overly powerful and prone to over-fitting the data (Horton, 1992).

10. Discussion

If only a few examples of a complex binding site are given, correspondingly little can be deduced about its statistical properties and even less about the true biological nature of the motif. The *E. coli* SD site provides a large data set for a non-trivial motif problem that should permit a relatively precise quantification and comparison of statistical and biological properties. For example, the presumed independence of matrix feature probabilities might not provide an adequate model of instance frequency or actual binding strength. The degree of correlation between instance frequency, motif strength, binding strength and biological effectiveness is of special interest, and the degree of confirmation or notable deviations are of equal interest.

The large amount of data also facilitates comparison of different motif representation languages. Surprisingly, despite its theoretical appeal, there is little evidence that matrix representation is actually better than other representations for characterizing biological motifs. Here, several types of motif representation were considered for characterizing the SD motif. As a general principle, the closer a representation language is to the actual recognition mechanism, the better the expected performance, and progressively relaxing the restrictions on a matrix representation allows progressively better characterization of the SD site. The prototypic *k*-mer (at least 5 of TAAGGAG) does a reasonable job, (signal/background ratio at 2000 USRs = 1.7) but IUPAC representation (eg [AGC]AGG[ATG]) is somewhat better (ratio = 2.8), and an unrestricted weight matrix representation appears to be the best (ratio = 3.6). However, too few restrictions on the representation language also leads to suboptimal performance as seen in the non-SD instances and over-fitting of the k-mer list motif.

Several new techniques for motif generation are described and compared to more standard approaches. At the 2000 USR comparison point, a variety of matrix approaches yield similar answers, indicating that the result is robust and that the different approaches are effective in finding near-optimum solutions. By these measures, when using a 4 x k matrix representation, about 1/4 of the USRs have no better than random SD sites. Whether this is as good as biological SD site detection remains to be determined. More biologically realistic models may well permit a better characterization of the SD motif.

It is remarkable that methods focussed on summarizing the frequency of SD sites (probability matrices) and those focussed on discriminating SD sites (weight matrices) sites should yield such similar results. It also seems remarkable that the probability matrices resulting from weight-matrix classification can frequently produce classification that is identical to the original weight matrix. Apparently evolutionary pressures lead to using a k-mer as an SD site in proportion to its discriminating power.

References

Benos, P.V, Lapedes, A. S., and Stormo, G. (2000). Is there a code for protein-DNA recognition. Probab(ilistical)ly. *BioEssays* **24**, 466-475.

Besemer, J., Lomsadze, A., Borodovsky, M.: (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, **29**, 2607-2618.

Fargo, D. C., Zhang, M., Gillham, N. W., Boynton, J. E. (1998). Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. *Mol. Gen. Genet.* **257**, 271-282.

Hampson, S. E., Kibler, D., and Baldi, P. (2002). Distribution Patterns of Over-Represented k-mers in Non-Coding Yeast DNA. *Bioinformatics* **18**, 513-528.

Hampson, S.E. and Volper, D.J. (1986). Linear Function Neurons: Structure and Training. *Biol. Cyber.* **53**, 203-217.

Horton, P. B., Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucl. Acids Res.* **20**, 4331-4338.

Hampson, S., Baldi, P., Kibler, D., and Sandmeyer, S. (2000). Analysis of yeast's ORFs upstream regions by parallel processing, microarrays, and computational methods. *ISMB2000*, 190-201.

Hampson, S. E., and Kibler, D. (1999). Minimum generalization via reflection: A fast linear threshold learner. *Machine Learning*, **37**, 51-73.

Hayes, W. S. and Borodovsky, M. (1998). Deriving Ribosomal Binding Site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction. *Pacific Symposium on Biocomputing* **3**, 279-290.

Karlin, S., Mrazek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *Jour. of Bateriology* **182**, 5238-5250.

Kibler, D., and Hampson, S. E. (2001). Learning weight matrices for identifying regulatory elements. *METMBS-2001*, 208-214.

Lewin, B. (2000). Genes VII. Oxford University Press.

Lukashin, A. V., Borodovsky, M. (1998). GeneMark.hmm new solutions for gene finding. *Nucleic Acids Research* **26**, 1107-1115.

O'Donnell, S. M., Janssen G. R. (2001). The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5' untranslated leader. *Jour. of Bacteriology* **183**, 1277-1283.

Sakai, H., Imamura, C., Osada, Y., Saito, R., Washio, T., Tomita, M. (2001). Correlation between Shine-Dalgarno sequence conservation and codon use of bacterial genes. *Jour. Mol. Evol.* **52**, 164-170.

Shine, J., Dalgarno, L. (1974). The 3'-terminal sequence of *E. coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *PNAS* **71**, 1342-1346.

Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E., Schneider, T. D. (2001). Anatomy of Escherichia coli Ribosome Binding Sites. *J. Mol. biol.* **313**, 215-228.

Stormo G. D., Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Bioinformatics* **14**, 691-9.

Stormo G. D. (2000). DNA binding sites: representation and discovery *Bioinformatics* **16**, 16-29.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *7th ISMB*, 262-271.

U. Keich and P. A. Pevzner. (2002). Finding motifs in the twilight zone. *Proceedings of the 6th international conference on computational molecular biology (RECOMB 2002).*