Characterizing the Shine-Dalgarno Motif: Probability Matrices and Weight Matrices

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

Abstract *Methods for identifying biologically significant k-mers by exhaustive evaluation ($k \leq 10$) are applied to the pooled Upstream Regions (USR) of all 4289 E. coli ORFs. Instances of the Shine-Dalgarno (SD) site are readily identified using these methods. Using these motif instances as starting points, two motif representations and training methods, probability and weight matrices, are applied to characterize the complete SD motif. Despite using different representations and objective functions, both methods yield approximately the same motif characterization, 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.*

Keywords: ribosome binding site, probability & weight matrices

1 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. 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 ([1]). 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 ([2]). Translation is possible without an SD site([3],[4]), 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 sequence of the 16s rRNA, the optimum SD sequence is TAAGGAG, and the degree of match is positively correlated with the rate of translation ([5]). The central subsequence AGGA is the most highly conserved part ([6]), but the entire 7-mer sequence shows some conservation.

The SD motif has a number of appealing features as a motif-learning test set, namely: 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 related SD sites, providing additional 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 computing k-mer binding strength, information that is not necessarily available when considering DNA-protein binding interactions; 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 increasingly 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. This last property means that it is difficult to make a simple categorical test as to whether a particular k-mer is an SD site or not. Instead we favor representations that a) are prototypically centered on the consensus sequence TAAGGAG, b) are highly-localized in the right region of the USR, and c) have a high signal-to-background ratio, as measured by the ratio of number of SD sites (the signal) identified in USRs versus those identified in scrambled sequences (the background). It is assumed that for two matrices that match the same number of sites in the real data, the one that matches the fewer random sites is preferable.

2 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. We have found that the ratio C0/C1 is quite effective in identifying over-represented, biologically significant k-mers ([7]). When applied to the 100 base USRs of 4289 ORFs in *E. coli*, C0/C1 identifies a number of highly over-represented strings, many with distinctive patterns of localization. Many of these appear to be SD sites since they are variations on the 7-mer TAAGGAG and are highly localized in a narrow hill centered about 13 bases upstream. 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, and based on a visual inspection of their localization in the USR, 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 motif shows high specificity for the SD region, it is assumed to be detecting real SD sites.

3 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 characterized in various ways, but here we focus on two methods: probability matrices and weight matrices.

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 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 (at least x of k) k-mer 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 monotonically 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.

If the base frequencies at each position are reasonably independent, feature frequency and feature importance for classification are quite close. However, feature frequency can be varied by changing instance frequency without changing the underlying importance. In practice the two are generally sufficiently close that a probability table can be used as a classifier or an estimate of binding energy ([8]), but the representations are not equivalent. Finding the optimum weight matrix suffers from the same search issues as probability matrices.

4 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. In addition, any motif evaluation metric incorporates certain, possibly inaccurate, assumption about the nature of the answer (eg minimum entropy, maximum over-representation, expectation maximization), so even the global optimum may not optimally characterize the biological process.

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 all of its M1 neighbors are also motif instances. With this assumption, plus the previous 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 1). 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, using the first column for example, the number of occurrences of the base string TAAGGAG is $.42 * 262 = 110$, since there are 110 exact matches. This probability table provides a possible characterization of the SD motif. Similar base strings yield similar

Table 1: Probabilities matrix based on M1 neighbors of TAAGGAG

	262	218	148	149	135	146	242
a	.27	.50	.74	.11	.11	.75	.31
t	.42	.03	.11	.08	.04	.07	.16
g	.13	.16	.06	.74	.81	.14	.45
c	.18	.31	.09	.07	.04	.04	.08

Table 2: Iterative probability matrix based on 3966 matches

a	.36	.43	.54	.08	.00	.72	.29
t	.33	.08	.16	.10	.01	.17	.19
g	.13	.22	.17	.76	.97	.00	.44
c	.18	.28	.13	.06	.02	.10	.08

M1 tables.

The M1 table based on the presumed prototypic sequence TAAGGAG is easy to compute, but is only an 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. 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 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 non-SD 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 2 in a few seconds. This table is based on 3996 matches and shows good localization in the USR. Starting from different initial k-mers similar to TAAGGAG and allowing either 0 or 1 mismatch in the initial probability table gave similar results.

There are at least 2000 ORFs with well-defined SD sites, so as point of comparison between representation, a threshold was set to classify approximately that many as positive instances. With this setting, the probability table produced 2102 matches in 2076 ORFs. In scrambled data (a first-order model) it produced 812 matches for a signal-to-background ratio of about 2.6. Beyond 3000 matches, the number of matches in the real data and the scrambled data are approximately the same, indicating that nearly 1/4 of the ORFs do not have SD sites in the 0 to 20 USR that are distinguishable from the background.

This result was achieved with a number of starting patterns, but there was some variability. One possible factor in the amount of variation in results is that the motif is being optimized 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 3), with a signal-to-background ratio of about 2.9. This is probably a slightly better characterization of the SD motif.

Table 3: Iterative probability matrix based on best 2000 matches

a	.37	.46	.83	.00	.00	.78	.32
t	.39	.00	.11	.00	.00	.11	.13
g	.12	.22	.00	1.0	1.0	.11	.50
c	.12	.32	.05	.00	.00	.00	.05

5 Weight Matrices

In previous work we extended the basic idea of k-mer over-representation ([9]) to groups of k-mers such as the M1 and M2 shells of a k-mer, IUPAC and weight matrix motif representations ([10],[11]). Over-representation can be calculated in different ways depending on the background model and the statistical method of computing over-representation. For example, as a background model, a k-mer's C0 in the data set might be compared to average k-mer frequency in the data set, the frequency of the k-mer in a different but related data set (eg the 20-40 base region), or the frequency of the k-mer is a scrambled version of the data set. Comparing to a scrambled version is the same as comparing to a first-order Markov Model of the data, which can be computed directly rather than actually scrambling the data. This suggests further possibilities using higher-order MMs 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. For convenience, we use z-scores here. 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 give similar answers, so the later method is used for speed.

For the SD problem, there is no obvious way to choose a region that is the same as the 0-20 region but without SD sites, so the simple expedient of using a first-order background model was employed. Higher-order models were tried but did not appear to be beneficial.

The goal is to adjust a weight matrix so as 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. Boundaries k-mers are those closest to the hyperplane defined by the current weight matrix and a threshold.

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 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 100 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 spend on multiple restarts. A single hill-climbing episode takes about 15 seconds.

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 4 shows the final weight matrix after being converted to probability matrix con-

Table 4: Weight matrix and resulting probability table based on 1302 matches

a	.25	.32	.45	.15	.16	.51	.25
t	.34	.00	.24	.15	.15	.13	.17
g	.19	.28	.07	.54	.55	.21	.43
c	.22	.40	.24	.15	.15	.15	.14
Threshold = 2.84							
a	.29	.42	.81	.00	.00	.97	.23
t	.47	.00	.11	.00	.00	.00	.07
g	.11	.15	.00	1.0	1.0	.02	.66
c	.13	.43	.08	.00	.00	.01	.04

straints and the resulting probability table over its positive instances. Using a multiplicative similarity rule, the probability table can be used to classify all 7-mers giving the same results as the weight matrix. In other cases an additive rule worked while multiplicative did not. There was no clear preference between multiplicative or additive scoring.

A potential problem with optimizing for over-representation is that the classifier with optimum over-representation need not have the desired coverage. 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 less over-represented than the other positive instances. This was in fact frequently observed to be the case. The final optimized weight matrices (all with very high z-scores) identified between 700 and 1300 SD instances when at least 3000 are known to exist. The identified SD instances show good localization, but greater coverage is desired. What is really wanted is the most over-represented set with a given coverage.

To address this issue, the algorithm was modified in an admittedly *ad hoc* way. In order to cover about 2000 ORFs, about 2200 k-mer matches are needed, based on experience with the previous algorithms. In order to encourage the algorithm to find the most over-represented set of about N matches, if the 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.

Weight matrix hill-climbing easily finds weighted combinations of features that are close to the desired coverage, and many similar local optima are produced (eg Table 5). This motif has a signal-to-background a ratio of 3.2 which is marginally better than the best probability tables. Using an additive rule, the probability table can classify k-mers the same as the weight matrix. The last row of this table gives the relative entropy of each position, demonstrating that all positions have information value except for the first one.

In general, weight matrix results are surprisingly similar 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, and in this case, a probability table over positive instances appears to be as effective for instance classification as a weight matrix. 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 ([12]), but such an authoritative classification is not currently available, and would be problematic given that SD instances appear to grade continuously into the random background.

6 Discussion

If only a few dozen 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 proper-

Table 5: Weight matrix and resulting probability table based on 2251 matches

a	.21	.31	.39	.18	.13	.43	.18
t	.31	.00	.21	.09	.18	.28	.25
g	.25	.36	.20	.54	.49	.15	.39
c	.23	.33	.20	.18	.20	.15	.18
Threshold = 2.54							
a	.27	.38	.76	.00	.00	.85	.19
t	.38	.00	.10	.00	.00	.11	.15
g	.20	.32	.07	.99	.98	.02	.60
c	.15	.30	.07	.00	.02	.02	.06
RE	0.0	0.3	0.4	1.3	1.3	0.6	0.3

ties. For example, the presumed independence of matrix feature probabilities is not true for simple (at least x of k) categories and might not provide an adequate model of actual binding strength. The degree of correlation between instance frequency, motif strength, actual binding strength and biological effectiveness is of special interest. While generally assumed to be the case, none of these properties are necessarily true, so the degree of confirmation or notable deviations are of equal interest. It is remarkable that methods focussed on summarizing the SD sites (probability matrices) and those focussed on discriminating SD sites (weight matrices) sites should yield similar results. Perhaps evolutionary pressures lead to using a k-mer as a SD site in proportion to its discriminating power.

References

[1] Lewin, B. Genes VII. Oxford University Press, 2000.

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

[3] Fargo, D. C., Zhang, M., Gillham, N. W., Boynton, J. E.: 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 (1998).

[4] O'Donnell, S. M., Janssen G. R. 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, 2001.

[5] Karlin, S., Mrazek, J. Predicted highly expressed genes of diverse prokaryotic genomes. Jour. of Bateriology 182: 5238-5250, 2000.

[6] Tompa, M. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. 7th ISMB, August 1999.

[7] Hampson, S. E., Kibler, D., and Baldi, P. Distribution Patterns of Over-Represented k-mers in Non-Coding Yeast DNA. Bioinformatics (in press)

[8] Stormo G. D., Fields, D. S. Specificity, free energy and information content in protein-DNA interactions. TIBS 23, March 1998.

[9] van Helden, J., Andre, B., and Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. JMB 281: 827-842, 1998.

[10] Hampson, S., Baldi, P., Kibler, D., and Sandmeyer, S. Analysis of yeast's ORFs upstream regions by parallel processing, microarrays, and computational methods. ISMB2000, August 2000.

[11] Kibler, D., and Hampson, S. E. Learning weight matrices for identifying regulatory elements. METMBS-2001. pp. 208-214, 2001.

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