



A probabilistic method for identifying start codons in bacterial genomes

Baris E. Suzek¹, Maria D. Ermolaeva², Mark Schreiber³ and Steven L. Salzberg^{1, 2,*}

¹Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA, ²The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD 20850, USA and ³Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand

Received on December 18, 2000; revised on April 12, 2001 and July 4, 2001; accepted on July 9, 2001

ABSTRACT

As the pace of genome sequencing has accelerated, the need for highly accurate gene prediction systems has grown. Computational systems for identifying genes in prokaryotic genomes have sensitivities of 98–99% or higher (Delcher *et al.*, *Nucleic Acids Res.*, 27, 4636–4641, 1999). These accuracy figures are calculated by comparing the locations of verified stop codons to the predictions. Determining the accuracy of start codon prediction is more problematic, however, due to the relatively small number of start sites that have been confirmed by independent, non-computational methods. Nonetheless, the accuracy of gene finders at predicting the exact gene boundaries at both the 5' and 3' ends of genes is of critical importance for microbial genome annotation, especially in light of the important signaling information that is sometimes found on the 5' end of a protein coding region. In this paper we propose a probabilistic method to improve the accuracy of gene identification systems at finding precise translation start sites. The new system, RBSfinder, is tested on a validated set of genes from *Escherichia coli*, for which it improves the accuracy of start site locations predicted by computational gene finding systems from the range 67–77% to 90% correct.

INTRODUCTION

At the initiation of protein synthesis, the ribosome binds to a region near the 5' end of the messenger RNA known as the Ribosome-Binding Site (RBS; Lewin, 1997). The RBS is typically located about 8–10 base pairs (bp) prior to the translation start site, and is usually characterized by a Shine–Dalgarno (SD) sequence pattern (Shine and Dalgarno, 1974), a 6 bp motif complementary to the 3' end of the 16S rRNA. This sequence varies among prokaryotes, but is generally highly conserved, a

reflection of the high conservation of 16S rRNA sequences (Mikonnen *et al.*, 1994). Our algorithm uses this essential property of prokaryotic genes to locate the RBS and the start site. The algorithm is designed to modify an initial set of gene predictions; therefore it must be run after a gene finder has been run. The current implementation is being used as a post-processor for the GLIMMER gene identification system (Delcher *et al.*, 1999), which prefers to locate the start site at the most upstream start codon for each gene. This implementation can also be used to process the output of the GeneMark bacterial gene finder (Lukashin and Borodovsky, 1998).

METHODS AND ALGORITHMS

The algorithm begins by finding a 'seed' sequence, which is then used by the algorithm to train a probabilistic model of ribosome-binding sites. It then uses the model to find RBSs in regions upstream of start codons, and these in turn are used to select the most likely start codon for a gene.

Finding the seed sequence

The training process begins with the identification of a seed sequence that forms a template for the model of the RBS. In most prokaryotes, the RBS is complementary to the 3' end of the 16S ribosomal RNA (rRNA). If the 16S rRNA of the organism is known, the simplest way to choose an organism-specific seed sequence is to use the 16S rRNA sequence.

Let R be the reverse complement of the last 15 bp of the 16S rRNA and let L be the length of the seed sequence (provided as an input to the algorithm). The input to the algorithm comprises a genome plus a complete list of gene locations; these gene locations can be supplied by an automated gene finder such as Glimmer (Salzberg *et al.*, 1998; Delcher *et al.*, 1999) or GeneMark.hmm (Lukashin and Borodovsky, 1998). To find the seed sequence, the algorithm uses the following simple procedure:

*To whom correspondence should be addressed.

- (1) generate the set S containing all subsequences of length L from R ;
- (2) for each gene, extract 30 bp adjacent to the start codon. Call these the upstream regions;
- (3) for each element in S , compute its frequency in the upstream regions;
- (4) select the element of S having the highest frequency as the seed.

If the 16S rRNA sequence is not known, and if no closely related organism's 16S rRNA is known, then one can use a *de novo* method for identifying motifs to generate a seed sequence. For example, Tompa (1999) developed an algorithm that finds conserved motifs upstream of the start codon. This method finds multiple motifs and assigns a probability (expressed as a z -score) to each one. The motif with the highest z -score can be used as the seed sequence for our algorithm.

Training the model

The RBS finding algorithm assumes that at least some of the start codons have been predicted correctly by the gene identification system; previous reports indicate that this assumption is correct for both Glimmer and GeneMark.hmm. The fact that average free-energy values fall sharply 10–20 bp upstream from start codons (Osada *et al.*, 1999) lends support to the theory that most RBSs reside within this region. Our method examines the sequences extending upstream of the start codons, using a window varying from 10–40 bp, and looks for conserved motifs in those regions. From each sequence in the upstream region, one subsequence is selected for the initial training set, determined by the following procedure.

Let Seq_i be a sequence within an upstream region beginning i bases from a start codon. $Seq_i[j]$ is the j th base of Seq_i and len_{seed} is the length of the seed sequence. For each Seq_i , calculate its similarity to the seed sequence Seq_{seed} as:

$$\text{Similarity}(Seq_{seed}, Seq_i) = \sum_{j=1}^{len_{seed}} Bpsim(Seq_{seed}[j], Seq_i[j])$$

where

$$Bpsim(Seq_{seed}[j], Seq_i[j]) = \begin{cases} 3 & Seq_{seed}[j] = Seq_i[j] \text{ and } Seq_{seed}[j] \in \{G, C\} \\ 2 & Seq_{seed}[j] = Seq_i[j] \text{ and } Seq_{seed}[j] \in \{A, T\} \\ 1.5 & Seq_{seed}[j] = A \text{ and } Seq_i[j] = G. \end{cases}$$

This simple similarity function uses the number of hydrogen bonds to weight each base pair, with a slightly lower weight given to $G-T$ bonds (Lewin, 1997).

The subsequence having maximum similarity to the seed is chosen for inclusion in the training set if it scores above a fixed threshold. The threshold is currently

Table 1. The accuracy of start codon predictions by RBSfinder on the Ecogene dataset (see Section **Results and discussion**) with different threshold values, using the seed sequence AGGAG and a window size of 15. The default threshold is 9. Accuracy was measured by using the algorithm applied as a post-processor for the Glimmer gene finding system

Threshold	Accuracy (%)
0	85.9
1	85.9
2	85.9
3	86.0
4	87.0
5	87.2
6	87.2
7	87.2
8	87.3
9	87.9
10	80.3
11	75.8
12	75.8
13	75.8

Table 2. The accuracy of start codon predictions made by RBSfinder for the Ecogene dataset using the most significant RBS motifs detected by Tompa's method. The most significant motifs for each length is underlined

Motif	Window size (bp)			
	15	20	25	30
<u>AGGA</u>	87	86	86	85
<u>AGGAG</u>	88	88	86	87
<u>CAGGAG</u>	81	81	81	79
<u>TAAGGAG</u>	72	72	72	72

computed as $3(len_{seed} - 2)$, which represents a relatively strong bond interaction between the sequence and the 3' end of the 16S rRNA. Table 1 reports the accuracy of start codon predictions on *Escherichia coli* genes using threshold values ranging from 0 to 15 (equivalent to 5 $G-C$ hydrogen bonds) for the seed sequence AGGAG and a window size of 15 bp. The data suggest that the algorithm is relatively insensitive to the choice of this threshold up to a value of 10; we used 9 as the default value.

Table 2 gives the accuracy of start codon predictions when Tompa's method is used to select the seed sequence, again using *E. coli* data. Each of the most significant motifs are a subsequence of the SD sequence (TAAGGAG). The method correctly identifies AGGAG as the ideal seed sequence, and the results also show that a seed length of 5 appears to generate the best overall performance.

Once all the training sequences are selected, they are put into a multiple alignment. We then construct a probabilistic model from the aligned training set as

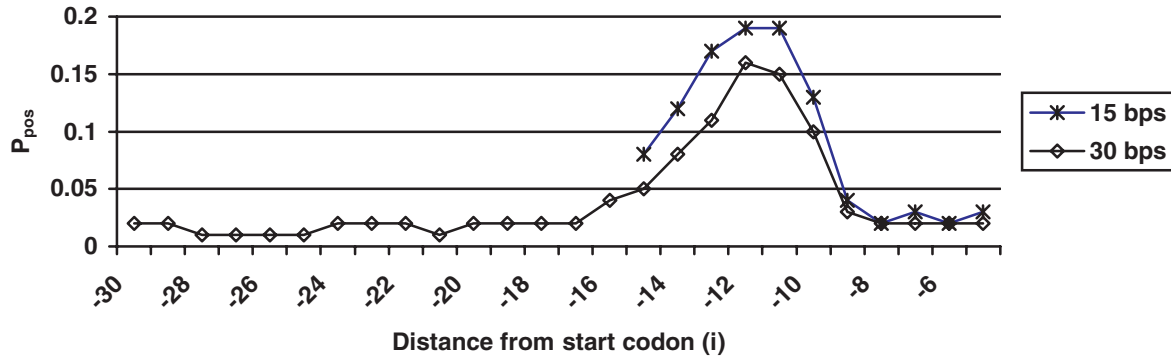


Fig. 1. Plot of probability of a training set sequence, $P_{\text{pos}}(i)$, occurring at varying distances upstream of the start codon, using the *E. coli* genome and start codon predictions made by Glimmer.

follows:

$$\begin{aligned}
 P(b, k) &= f(b, k)/N \quad \text{for } b \in \{A, T, G, C\} \\
 &\quad \text{and } \text{len}_{\text{seed}} \geq k \geq 1 \\
 P_{\text{pos}}(i) &= f_{\text{pos}}(i)/N \quad \text{for } W \geq i \geq \text{len}_{\text{seed}} \quad (1)
 \end{aligned}$$

where $P(b, k)$ and $f(b, k)$ are the probability and frequency (respectively) of base b occurring in position k in the training set; N is the number of sequences in the training set; $P_{\text{pos}}(i)$ and $f_{\text{pos}}(i)$ are the probability and frequency (respectively) of a training set sequence occurring i nucleotides upstream of the start codon; and W is the window size. A plot of $P_{\text{pos}}(i)$, versus distance from the start codon reveals that most of the sequences selected for training are located in a region 9–15 bp upstream of the start codon (Figure 1).

Finding ribosome binding sites

In order to find a ribosome-binding site, the algorithm searches all upstream regions using the following iterative procedure.

- (1) For each subsequence Seq_i of length len_{seed} within an upstream region, calculate a combined score C_i using the formula:

$$C_i = P_{\text{pos}}(i) \times \text{Score}_i \quad (2)$$

where

$$\begin{aligned}
 \text{Score}_i &= \prod_{k=1}^{\text{len}_{\text{seed}}} P(\text{Seq}_i[k], k) \quad \text{for } \text{Seq}_i[k] \\
 &\in A, T, G, C \quad (3)
 \end{aligned}$$

where $\text{Seq}_i[k]$ is the k th base of the subsequence Seq_i . The Score_i function is a standard position weight matrix (Claverie and Audic, 1996).

- (2) For each gene, select the sequence $\text{Seq}_{\text{highest}}$ from the upstream region with the highest combined score C_{highest} as the candidate RBS for that gene.
- (3) If $\text{Seq}_{\text{highest}}$ has a score greater than a computed threshold (described below), then use it as the ribosome-binding site.

Relocating start codons

The system investigates all possible ribosome-binding sites around each predicted start codon using the procedure described in the previous section. If better RBSs are found either upstream or downstream of the originally predicted start, then the system moves the start codon accordingly. Upstream and downstream start codons must be in the same reading frame in order to be considered. After finding the RBSs for all alternative start sites, the new start codon is selected based on the RBS score in equation (2) and the following rules:

- (1) The start codon ATG is favored over GTG or TTG, even if the ATG has a lower-scoring RBS (all sites must score above the threshold).
- (2) GTG is favored over any TTG.
- (3) If the start codons are the same, the site with the higher-scoring RBS is chosen.

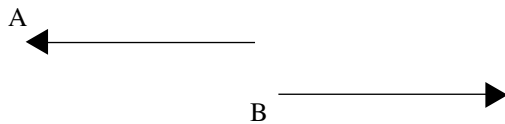
These rules improved the accuracy by 1% for the Eco-gene dataset (see below) compared to the alternative of selecting the new start codon solely based on the combined score of the ribosome-binding sites.

The threshold RBS score used for the original start codon and for upstream alternatives is computed as $\text{Threshold}_{\text{up}} = (\text{Score}_{\text{max}} + \text{Score}_{\text{min}})/2 \times \min(P_{\text{pos}})$, where $\text{Score}_{\text{max}}$ and $\text{Score}_{\text{min}}$ are the scores of the best and worst initial RBSs (respectively) from the training set, computed according to equation (3). $\min(P_{\text{pos}})$ is the minimum $P_{\text{pos}}(i)$ from equation (1); i.e. the probability

Table 3. Variation in the accuracy of start codon predictions made by RBSfinder as genes are allowed to overlap. These results used the seed sequence AGGAG and a window size of 15 for the Ecogene test set

Maximum overlap length allowed (bp)	0	10	20	30	40	50	60
Accuracy (%)	87.92	88.06	88.06	88.20	88.20	88.20	88.20

associated with the least likely place for an RBS to occur in the training set. Note that the upstream relocation of the start codon might lead to overlapping genes, especially where intergenic regions are small. This issue is resolved as follows: suppose two genes A and B are transcribed in opposite directions, away from each other, such as:



In this case, the algorithm checks for an RBS associated with gene B prior to adjusting the start location of A. If B already has a good RBS, the algorithm will consider alternative start codons for A only in the intergenic region between A and B. If B does not have a strong RBS, then the algorithm searches within B, as long as the reading frame of A can be extended further.

If A and B are transcribed in the same direction, with B following A, the algorithm will only check start codons for B as far back as the stop codon of A. If the initial start codon location of B overlaps the end of A, the algorithm only considers alternative starts in the downstream direction for B. The algorithm prefers not to allow overlapping genes, but it does permit them. Table 3 shows the effect of allowing overlapping on the accuracy of start codon predictions for *E. coli*; a small increase in accuracy results when genes are allowed to overlap by as much as 30 bp.

The threshold score used for RBSs associated with downstream start codons is computed as:

$$\text{Threshold}_{\text{down}}(k) = \text{Threshold}_{\text{up}} + \left((C_{\text{max}} - \text{Threshold}_{\text{up}}) \times \frac{\text{moveDist}}{\text{moveMax}} \right)$$

where

$$C_{\text{max}} = \text{Score}_{\text{highest}} \times \max(P_{\text{pos}})$$

where moveDist is the distance that the start codon would have to move to end up in location k , moveMax is a function that determines the maximum distance that the system is allowed to move a start codon in the downstream direction, and $\text{Score}_{\text{highest}}$ the score of the best RBS from

Table 4. The accuracy of start codon predictions made by RBSfinder for the Ecogene dataset with different threshold functions, using the seed sequence AGGAG and a window size of 15 where $\text{Score}_{\text{avg}} = (\text{Score}_{\text{lowest}} + \text{Score}_{\text{highest}})/2$ and $P_{\text{avg}} = (\max(P_{\text{pos}}) + \min(P_{\text{pos}}))/2$

Threshold function	Accuracy (%)
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{highest}} \times \min(P_{\text{pos}})$	76.3
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{lowest}} \times \min(P_{\text{pos}})$	86.6
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{highest}} \times \max(P_{\text{pos}})$	70.5
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{lowest}} \times \max(P_{\text{pos}})$	87.6
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{lowest}} \times P_{\text{avg}}$	81.3
$\text{Threshold}_{\text{up}} = \text{Score}_{\text{avg}} \times \min(P_{\text{pos}})$ (default)	87.9
$\text{Threshold}_{\text{down}} = \text{Score}_{\text{avg}} \times P_{\text{avg}}$	84.5
$\text{Threshold}_{\text{down}} = \text{Threshold}_{\text{up}}$	82.2
$\text{Threshold}_{\text{down}} = \text{Threshold}_{\text{up}} = \text{Score}_{\text{highest}} \times \min(P_{\text{pos}})$	75.8
$\text{Threshold}_{\text{down}} = \text{Threshold}_{\text{up}} = \text{Score}_{\text{lowest}} \times \min(P_{\text{pos}})$	80.9
$\text{Threshold}_{\text{down}} = \text{Threshold}_{\text{up}} = \text{Score}_{\text{highest}} \times \max(P_{\text{pos}})$	70.5
$\text{Threshold}_{\text{down}} = \text{Threshold}_{\text{up}} = \text{Score}_{\text{lowest}} \times \max(P_{\text{pos}})$	81.3

equation (3). $\text{Max}(P_{\text{pos}})$ is the maximum $P_{\text{pos}}(i)$ from equation (1); i.e. the probability associated with the most likely position of an RBS in the training set. This threshold function is designed to ensure that moving the start site further downstream will require a higher-scoring RBS (i.e. the threshold increases with the distance of the move). The parameter moveMax is designed to prevent the system from shortening a gene too drastically when it moves the start site downstream; in our experiments moveMax was set to 35% of a gene's length for genes longer than 300 bp, and 20% of gene length otherwise. If the algorithm does not find a ribosome-binding site either in the upstream or downstream directions, it leaves the start codon in its original location. Table 4 shows the variation in accuracy of start codon predictions as a function of the choice of threshold function. Based on the results in this table, the threshold functions used by the system were chosen to be the functions described above.

RESULTS AND DISCUSSION

The difficulty with evaluating any method for finding start sites (and ribosome binding sites) is the relatively small number of genes for which the start sites have been verified in a laboratory. The vast majority of genes in bacteria genomes, including those in *E. coli*, have had their start sites predicted computationally rather than experimentally. In order to test the RBSfinder algorithm, we used only experimentally verified genes for *E. coli*, which contains by far the largest number of genes with validated starts among all bacteria, although even here the number is only a small fraction of the total complement of 4288 protein-coding genes (Blattner et al., 1997). The Ecogene database (Rudd, 2000) contains 717 proteins

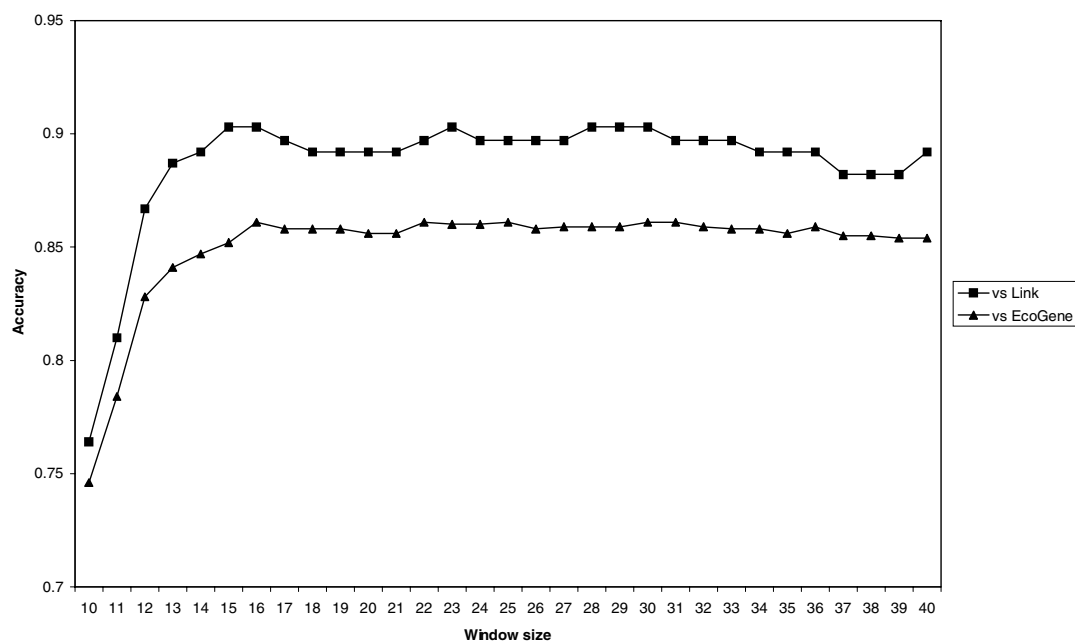


Fig. 2. Accuracy of start codon predictions by RBSfinder as a post-processor for Glimmer gene predictions.

that have been verified by N-terminal protein sequencing; these were extracted and used as the test set for our algorithm. The 'Link' dataset (Link *et al.*, 1997) of 195 N-terminally verified genes was also used. This subset of the Ecogene dataset contains only genes that either have a processed leader sequence of a single amino acid or do not have a processed leader sequence and therefore do not require an estimation of the correct start site based on a putative leader sequence. Thus we assume the Link subset is slightly more accurate than the full dataset.

The Glimmer 2.0 system (Delcher *et al.*, 1999) correctly predicts the startcodon 472 times (66%) in the Ecogene dataset and 133 times (68%) in the Link dataset. After post-processing by RBSfinder, the accuracy of start codon predictions increased to 88 and 92% on Ecogene and Link respectively. Accuracy varied according to the window size chosen (Figure 2), with the most accurate results using a window size of 16–17 bp. The GeneMark.hmm system (Lukashin and Borodovsky, 1998) identifies the correct start codon 550 times (77%) in the Ecogene dataset. After post-processing by RBSfinder, this accuracy improves to 640 correct start codons (90%). Thus in both cases, the use of RBSfinder produces substantial improvements in the accuracy of start codon prediction, raising the accuracy of both systems to near 90%.

The design of the RBSfinder algorithm allows for it to adjust start codon locations repeatedly if the algorithm is run more than once. It quickly converges, however, usually in just 4–5 iterations. These successive applications allows

the algorithm to improve the accuracy of start codon predictions further, as shown in Figure 5. Thus the recommended mode of running the system is to run it several times in succession; the entire process takes just a few minutes. The code is freely available for download from <http://www.tigr.org/softlab>.

Because the test data does not contain validated RBSs (only validated start sites), we can at best assume the RBS is correct when the system finds an RBS near a correct start codon. To compute the algorithm's specificity, genes that were predicted to have an RBS and that also had a correct start codon prediction were considered to be True Positives (TP). Those genes that had an RBS prediction but an incorrect start codon prediction were considered to be False Positives (FP). Specificity was then calculated as $TP/(TP + FP)$. Since the thresholds for selecting an RBS upstream, downstream and at the original start site are all different, we calculated specificity separately for each of these three types of predictions. Sensitivity was calculated as $TP/(\text{Size of validated set})$.

As might be expected, sensitivity generally increased as the window size was increased (Figure 3). Sensitivity must, however, be considered in light of specificity, which was generally maximal at window sizes of between 11 and 18 bp (Figure 4). Reassuringly this coincides with the size of the upstream ribosome footprint. Using a window size of 16 bp, for example, the system finds approximately 85% of ribosome binding sites (sensitivity) with a specificity ranging from 80% (for start codons left

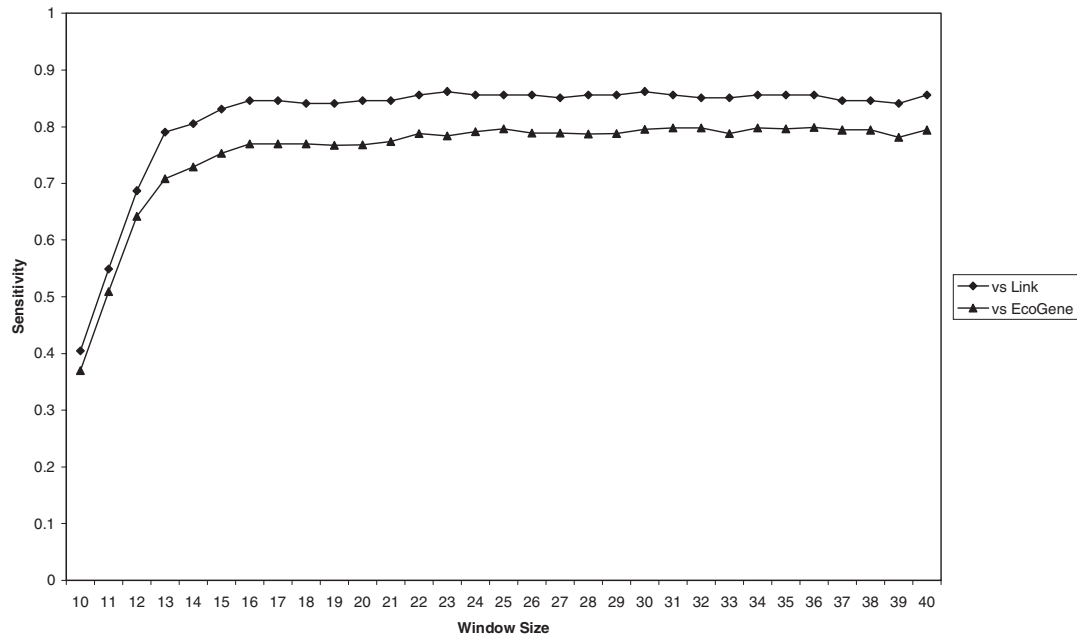


Fig. 3. Sensitivity of RBSfinder to find ribosome binding sites for Glimmer gene predictions.

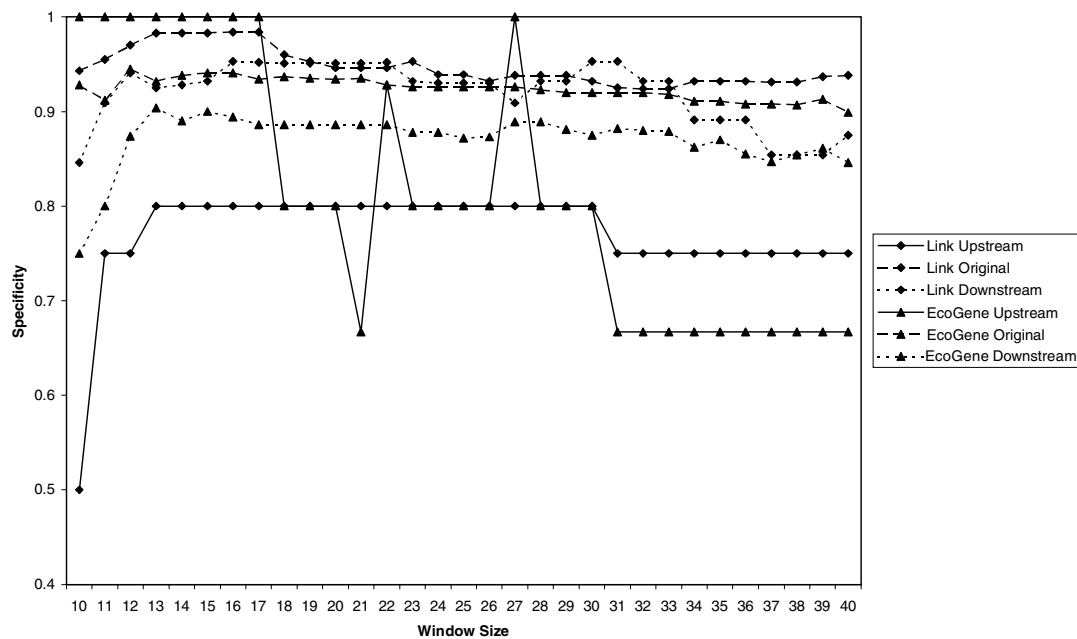


Fig. 4. Specificity of RBSfinder for identifying ribosome binding sites in various *E. coli* datasets.

in their original location) to near 100% (for start codons moved upstream).

Most bacterial genomes are not as well studied as *E. coli*, and extensive experimental data for them is not available. Nonetheless, we would like to estimate the

accuracy of the RBS finder on other organisms. To do this, we need a method for estimation that does not rely on experimental data. In order to make this estimate, we compared the number of RBSs found at the original start sites (using Glimmer's predictions) with the number of

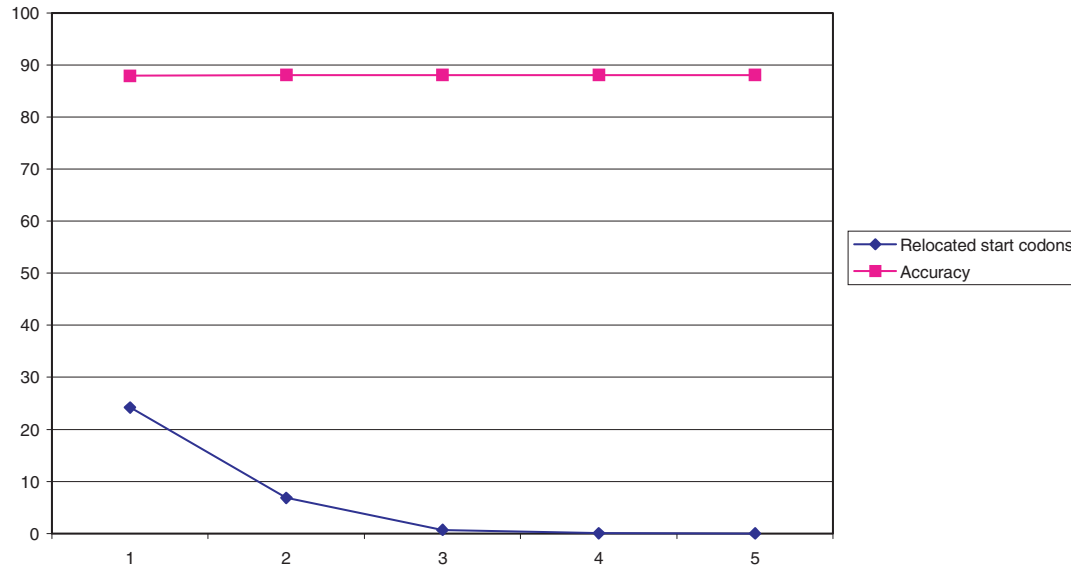


Fig. 5. The accuracy of start codon predictions on the Ecogene dataset using RBSfinder as a post-processor for Glimmer. Also shown is the number of start codons relocated after each successive application of RBSfinder.

RBS found at the same locations in a pseudo-random DNA sequence with the same GC-content. The model was trained on real sequence data in both conditions. We then used the frequency of RBSs found in random sequence as the background FP rate. Let $RBS(\text{real})$ represent the frequency of finding RBSs in a real sequence, and let $RBS(\text{random})$ represent the same quantity for a random sequence, then:

$$k = \frac{RBS(\text{real}) - RBS(\text{random})}{RBS(\text{real})} \cdot 100\%$$

where $k \approx 100\%$ means that all (or almost all) predictions of RBSs are correct. Values under 50% indicate that more than half of the putative RBSs identified computationally are likely to be false; thus it would be better to leave the original gene finder's predictions alone. The value k can be considered a theoretical estimate of specificity; note that the estimate based on experimental data is higher than the theoretical estimate of 63% found here for *E. coli*. One possible explanation for this difference is that evolutionary pressure may eliminate 'random' (non-functional) ribosome binding sites from the region just upstream of a start codon. This value might best be considered then as a conservative estimate of specificity. Sensitivity was estimated as the percentage of genes for which a ribosome binding site was predicted.

The values of k (specificity) and sensitivity for a small sample of microbial genomes are shown in Table 5. These estimates give a rough guide to whether or not RBSfinder will be a useful method for adjusting the

Table 5. Theoretical estimates of specificity (k) and sensitivity of RBSfinder, for different microbial genomes

Species	15 bp window		30 bp window	
	Specificity, %	Sensitivity, %	Specificity, %	Sensitivity, %
<i>B.subtilis</i>	69	78	54	84
<i>A.aeolicus</i>	69	58	56	77
<i>E. coli</i>	63	66	52	74
<i>H.influenzae</i>	60	82	53	82
<i>A.fulgidus</i>	53	61	43	69
<i>M.genitalium</i>	27	29	34	67

start site predictions for a genome. For example, the data in Table 5 shows that the program (with AGGAG used as the seed sequence) is expected to be more effective for *Bacillus subtilis*, *Aquifex aeolicus*, *E. coli*, *Haemophilus influenzae*, and *Archaeoglobus fulgidus* than it is for *Mycoplasma genitalium*. Table 5 also shows that the specificity of the program usually decreases when window size is increased from 15 to 30 bp, consistent with the results from the *E. coli* experimental data.

ACKNOWLEDGEMENTS

S.L.S. and M.D.E. were supported in part by grants IRI-9902923 and KDI-9988088 from the National Science Foundation and grant R01-LM06845 from the National Institutes of Health. M.S. was supported by a Targeted PhD Scholarship from Otago University obtained on his behalf by Christopher Brown.

REFERENCES

- Blattner,F.R., Plunkett,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
- Claverie,J.-M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrices. *Comput. Appl. Biosci.*, **12**, 431–440.
- Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
- Fleischmann,R.D., Adams,M., White,O., Clayton,R., Kirkness,E., Kerlavage,A., Bult,C., Tomb,J.-F., Dougherty,B., Merrick,J., McKenney,K., Sutton,G., FitzHugh,W., Fields,C., Gocayne,J., Scott,J., Shirley,R., Liu,L.-I., Glodek,A., Kelley,J., Weidman,J., Phillips,C., Spriggs,T., Hedblom,E., Cotton,M., Utterback,T., Hanna,M., Nguyen,D., Saudek,D., Brandon,R., Fine,L., Fritchman,J., Fuhrmann,J., Geoghagen,N., Gnehm,C., McDonald,L., Small,K., Fraser,C., Smith,H. and Venter,J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Lewin,B. (1997) *Genes VI*. Oxford University Press, Oxford.
- Link,A.J., Robison,K. and Church,G.M. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli*. *Electrophoresis*, **18**, 1259–1313.
- Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
- Mikonnen,M., Vuoristo,J. and Alatossava,T. (1994) Ribosome binding site consensus sequence of *Lactobacillus delbrueckii*. *FEMS Microbiol. Lett.*, **116**, 315–320.
- Osada,Y., Saito,R. and Tomita,M. (1999) Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics*, **15**, 578–581.
- Rudd,K.E. (2000) Ecogene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
- Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
- Shine,J. and Dalgarno,L. (1974) The 3'-terminal sequence of *E. coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.
- Sprenghart,M.L. and Porter,A.G. (1997) MicroReview: functional importance of RNA interactions in selection of translation initiation codons. *Mol. Microbiol.*, **24**, 19–28.
- Tomba,M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Seventh International Conference on Intelligent Systems for Molecular Biology*. Heidelberg, Germany, pp. 262–271.