 Nucleic Acids Research, 1999, Vol. 27, No. 17 **3577–3582**

Bacterial start site prediction

Sridhar S. Hannenhalli*, William S. Hayes, Artemis G. Hatzigeorgiou[1] and James W. Fickett

Bioinformatics, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, PA 19406, USA and [1]Synaptic Ltd, Science and Technology Park of Crete, PO Box 1447, Voutes Heraklion, 71110 Greece

Received April 12, 1999; Revised June 25, 1999; Accepted July 8, 1999

ABSTRACT

With the growing number of completely sequenced bacterial genes, accurate gene prediction in bacterial genomes remains an important problem. Although the existing tools predict genes in bacterial genomes with high overall accuracy, their ability to pinpoint the translation start site remains unsatisfactory. In this paper, we present a novel approach to bacterial start site prediction that takes into account multiple features of a potential start site, viz., ribosome binding site (RBS) binding energy, distance of the RBS from the start codon, distance from the beginning of the maximal ORF to the start codon, the start codon itself and the coding/non-coding potential around the start site. Mixed integer programing was used to optimize the discriminatory system. The accuracy of this approach is up to 90%, compared to 70%, using the most common tools in fully automated mode (that is, without expert human post-processing of results). The approach is evaluated using *Bacillus subtilis*, *Escherichia coli* and *Pyrococcus furiosus*. These three genomes cover a broad spectrum of bacterial genomes, since *B.subtilis* is a Gram-positive bacterium, *E.coli* is a Gram-negative bacterium and *P.furiosus* is an archaebacterium. A significant problem is generating a set of 'true' start sites for algorithm training, in the absence of experimental work. We found that sequence conservation between *P.furiosus* and the related *Pyrococcus horikoshii* clearly delimited the gene start in many cases, providing a sufficient training set.

INTRODUCTION

As of March 1999, 20 bacterial genome sequences have been published and sequencing of many more is in progress. The complete list of these sequences is available from the public database (http://www.tigr.org). With the growing number of completely sequenced bacterial genes, accurate gene prediction in bacterial genomes remains an important problem. Significant progress has been made in the past few years in developing computational tools for gene prediction, GeneMark (1,2) and GLIMMER (3) being the most widely used tools for bacterial gene prediction *in silico*.

Although the existing tools predict the genes in bacterial genomes with high overall accuracy, their ability to pinpoint the translation start site remains unsatisfactory. In order to analyze the putative protein product of a gene, it is valuable to know as accurately as possible the translation initiation site. The two main sources of evidence used in finding bacterial genes are long open reading frames (ORFs) and some form of statistical regularity from codon usage bias, typically measured in a window of ~100 bp. While both kinds of evidence help a great deal in providing rough gene locations, neither one helps the investigator very much in choosing between alternative start codons near the beginning of an ORF.

The so-called Shine–Dalgarno consensus sequence (Shine and Dalgarno, 1974) is often used to search by eye for the ribosome binding site (RBS), but there are a number of more reliable methods (reviewed in 4). Stormo and colleagues (5) present one of the pioneering works in the computational characterization of translation start sites in prokaryotes. Schurr and colleagues (6) developed an algorithm for calculating the optimal binding energy between the 16S rRNA of *Escherichia coli* and the region upstream of a potential initiation codon, allowing internal loops and bulges, and showed a difference in the binding energy distribution for regions upstream of true initiation codons and spurious, gene-internal, ATG codons.

The study of Schurr *et al*. suggests that a practical gene start prediction method might be made on the basis of an optimal binding energy calculation. High accuracy on the basis of the RBS might be difficult in *E.coli*, where the RBS pattern is rather weak. However, in clostridial Gram-positive bacteria, in particular *Bacillus subtilis* and *Staphylococcus aureus*, the predicted energy at the potential RBS tends to be much stronger, perhaps because in these organisms the translation initiation complex is missing ribosomal protein S1, thought to help in melting inhibitory secondary structures in the mRNA (reviewed in 7). An energy-based algorithm to locate the RBS in these organisms was implemented in (8).

M. Borodovsky's group pioneered improved start site localization in bacterial gene finders, culminating in (9,10). The final synthesis in (9) uses such factors as the start codon score, RBS score, downstream box score, pre-start signal score and post-start signal score, all based on similarity to profiles generated from a training set.

Frishman and colleagues further explored the idea of using diverse evidence (11). The evidence used for gene recognition was coding potential and the evidence for start site prediction was RBS score which included a profile based score as well as

*To whom correspondence should be addressed. Tel: +1 610 270 6270; Fax: +1 610 270 5580; Email: hannes00@mh.vs.sbphrd.com

a score depending on the distance of the RBS to the start codon.

In this paper we present one more approach to bacterial start site prediction that takes into account multiple features of a potential start site, viz., RBS binding energy, distance of the RBS from the start codon, distance from the beginning of the maximal ORF to the start codon, the start codon itself and the coding/non-coding potential around the start site. There is a biological rationale supporting each of these factors as discussed in Materials and Methods.

For the developer, the main innovation here is that the discrimination problem is taken to be the same as the one faced by the ribosome: to choose one start codon, from among the first few potential starts in the ORF, at which to initiate translation. A true optimization method, mixed integer programing (MIP) is used to derive a discriminatory model appropriate to this formulation of the problem. In cases where different translation start sites are thought to be used on different occasions, the final optimized scoring of start sites can still be used to rank the possibilities.

For the user, the main innovations are: (i) the energy function used to evaluate a potential RBS is biologically motivated and allows for gapped alignments between the 16S rRNA and the RBS (important for at least some genes); (ii) the biological preference for start codons early in the ORF is incorporated in the algorithm in a very natural way; (iii) in each of three widely divergent species (*B.subtilis*, a Gram-positive bacterium; *E.coli*, a Gram-negative bacterium and *Pyrococcus furiosus*, an archaebacterium), a set of true start sites was hand selected based on the best available evidence, and the algorithm cross validated in each species; and (iv) no hand-tuning of the algorithm is required.

Most computational techniques for finding genes tend to be rather organism-specific and require a large training set of known genes to parameterize them for a new genome (4,12). This presents, of course, a problem with new genomes appearing very rapidly and with few genes known with any certainty in some of the new genome sequences. Thus increasing emphasis is being placed, among algorithm developers, on methods that can discover patterns *de novo*, in completely unannotated genomes. For example, both GeneMark-Genesis (13) and GLIMMER (3) use long ORFs to derive models of coding sequence. Grosse and colleagues have developed a coding region statistic based on mutual information that is organism independent and performs about as well as most codon usage statistics.

Searching for genes using potential homologs from related species, formalized in the Procrustes algorithm (14) and Critica, is also an important means of looking for genes without depending too much on peculiarities of the species at hand. A significant problem is generating a set of 'true' start sites for algorithm training, in the absence of experimental work. We found that sequence conservation between *P.furiosus* and the related *Pyrococcus horikoshii* clearly delimited the gene start in many cases, providing a sufficient training set in at least this case.

MATERIALS AND METHODS

The novel approach was tested on *B.subtilis*, *E.coli* and *P.furiosus*. For *B.subtilis*, the sequence was taken from GenBank (accession number AL009126) (15). The genes in the GenBank annotation with names starting with a 'non-y' character were taken to have correctly indicated starts. These are individually verified for the start sites as per the naming convention (A.Danchin, personal communication). This set included 1246 genes. *Escherichia coli* sequence was taken from GenBank (accession number U00096) (16). A set of 184 high confidence *E.coli* N-terminals were taken from (17), where N-terminals were derived from direct protein sequencing. From the original set of 223 N-terminals given in that study, several were removed since they were predicted using observed protein sequence where Edman sequencing was blocked or initiated at a significant distance from the likely N-terminal. The sequence for *P.furiosus* was provided by Bob Weiss at the Utah Genome Center (http://www.genome.utah.edu), and a set of start sites was extracted via homology with *P.horikoshii* (GenBank accession numbers AP000001–7) ORFs.

Pyrococcus furiosus maximal ORFs were first aligned against several archaeal genomes, including *Archaeoglobus fulgidus*, *Methanococcus jannaschii*, *Methanobacterium thermoautotropicum* and *P.horikoshii*, using BLAST (18). (These alignments were used to derive the training set but were not, of course, used in the final prediction algorithm.) Only the alignments with *P.horikoshii* that showed strong homology and a crisp drop in conservation could be used for start site prediction. The pairs of genes with >80% identity over 200 bp were realigned at the amino acid level using FASTA3 (19). The alignments included 33 amino acids upstream of the maximal ORF start in both *P.furiosus* and *P.horikoshii*. Next, each of the alignments was examined individually, via a plot of percent amino acid identity in a moving 15-residue window. Those with no ambiguity were chosen for the test/train set. A typical such case, of homology between related *P.furiosus* and *P.horikoshii* genes, is shown in Scheme 1.

The final *P.furiosus* set contains 240 genes including 15 experimentally characterized genes from the literature.

For each genome, the following analysis was done. For each gene g (uniquely identified by the stop site locus e_g), a complete list S_g of potential start sites was generated which obviously contains the 'true' start s_g. This list contains all in-frame start codons positioned downstream of the closest in-frame stop codon upstream of e_g.

For each potential start site s belonging to S_g, we compute five parameters: (i) the distance of the start codon from the maximal start in S_g; (ii) the start codon; (iii) the binding energy between the 3' end of 16S rRNA and the region of the genome immediately upstream of the putative start site; (iv) the length of the gap between the end of the RBS and the start codon; and (v) the score of the start codon based on coding potential of the regions upstream and downstream of the start codon.

All the above factors are biologically motivated. The rationale behind factor 1 is the fact that the 5' of the gene is available to the translation machinery prior to the 3' end. Factors 2, 3 and 4 play a role in the stability of the translation initiation complex, and factor 5, which is a function of codon usage around the start site, captures both the well-known fact of codon bias as well as, perhaps, some aspect of the role of RNA structure in start site specificity.

As an illustration, a gene g starting at position 88 and ending at position 906 of the genome will be converted into a list of potential starts and associated parameters as shown in Table 1.

```
>>PH0224_198383_199954                                    (557 aa)
 initn: 3081 init1: 3081 opt: 3081
Smith-Waterman score: 3081;  81.836% identity in 523 aa overlap

                    10        20        30        40        50        60
furiosus    MXSLKTVFLNYRLLTVTLKXCSPSYLWWXKMVHWADYMAEKIIKERGEKEEYVVESGITP
                                          :::::::.:.:::.:::::::.:::::::::
horikoshii QRNYYMLQWRNKLLFLNTXKHNINSKFGDYMVHWADYIADKIIRERGEKEKYVVESGITP
                    10        20        30        40        50        60

                    70        80        90       100       110       120
furiosus    SGYVHVGNFRELFTAYIVGHALRDRGYNVRHIHMWDDYDRFRKVPKNVPQEWEEYLGMPV
            :::::::::::::::::::::::.::::::::::::::::::.:::::::.:::::::.
horikoshii SGYVHVGNFRELFTAYIVGHALRDKGYEVRHIHMWDDYDRFRKVPRNVPQEWKDYLGMPI
                    70        80        90       100       110       120
```

Scheme 1.

Table 1. Set of potential starts and associated parameters for the gene ending at position 906

Start	Stop	Offset	Codon	Energy	Gap	Coding potential
73	906	0	atg	–12.700	–2	0.900
88	906	15	gtg	–6.400	9	0.500
373	906	300	ttg	–4.600	–1	0.000
499	906	426	ttg	–2.400	4	0.000
619	906	546	ttg	–3.800	7	0.000
703	906	630	ttg	–8.000	7	0.000
757	906	684	atg	–5.500	5	0.000
784	906	711	atg	–6.200	2	0.000

The binding energy was computed using a dynamic programing algorithm developed in (8) and later updated by S. Hannenhalli to reflect the energy and loop parameters in (20). The start score based on coding potentials was computed using GeneMark. This is a crude measure of how likely the start is, based upon it being at the boundary of non-coding and coding sequence. The calculation for the start score is given as

GeneMark start score = $P^{noncod} * P^{cod}$,

where P^{noncod} refers to the probability of non-coding, based upon the Markov model used by GeneMark, for one window width (which is a default of 96 nt in GeneMark) upstream of the start codon, and P^{cod} is the probability of coding for one window width downstream of the start codon in the frame defined by the start codon.

The values corresponding to each of these factors are converted into a log-odds estimate. Under this measure, the support by a parameter P with value p is measured as

$LogOdd(P = p) = \log[(\textit{frequency of } P = p \textit{ in true set})/(\textit{frequency of } P = p \textit{ overall})]$

Represent the five log-odd scores for the five parameters for the potential start s as $rank_s$, $codon_s$, $energy_s$, $spacing_s$ and $CodPot_s$ respectively.

Define a function

$Score_s = w_r * rank_s + w_c * codon_s + w_e * energy_s + w_s * spacing_s + w_C * CodPot_s$

where w's are relative weights of the parameters. For a given set of weights, the potential start s maximizing $Score_s$ is chosen as the start of the gene.

For a set of 'true' genes, the set of all potential start sites other than the true one make a negative data set. The weights of the parameters are trained on the training set so as to maximize the number of true positives. Most of the problems that arise in this field, requiring some discriminatory approach, have one pool of positive data and one pool of negative data and the attempt is made to discriminate between these two pools. What makes this problem different is that the decision is to be made within each set (corresponding to potential starts for a gene) where it is known *a priori* that there is exactly one true site in each set. Fortunately, MIP (21) models this problem quite well. For a recent application of MIP to sequence analysis see (22). In the following, we describe our MIP models to compute an optimal set of weights.

Let n be the number of genes. Let m be the number of potential starts (considered) for each gene. In all the genomes studied, the true start for a gene was among the first five (5' to 3') potential starts in >98% of the cases. To make it precise, among 1246 true starts in *B.subtilis*, 785 were at rank 1, 280 at rank 2, 105 at rank 3, 38 at rank 4, 20 at rank 5. For all practical purposes, m could be taken as 5. We limit m only to simplify the presentation and the implementation of the model, since the constraints can be expressed concisely for uniform m.

For the i^{th} gene g^i, the potential starts are represented as $s^{i,j}$, $1 \le j \le m$. Without loss of generality, we assume that $s^{i,1}$ is the 'real' start and $s^{i,j}$, $2 \le j \le m$ are the 'false' starts. For a given (potential) start $s^{i,j}$, the five log-odd values corresponding to the five factors are denoted by $v^{i,j}_k$, $1 \le k \le 5$, respectively. Also, let w_k, $1 \le k \le 5$ be the five weights to be computed. We fix $w_1 = 1$ to normalize the weights. Let M be a 'sufficiently large' integer. In Figure 1, we present the first MIP model that we used to compute the weights.

This model minimizes the number of genes in which a violation occurs, i.e., the score of the 'true' start is not the maximum. This is because f^i must be 1 for gene i as long as the 'true' start does not have the maximum score. This model does not capture

$$GIVEN : v_k^{i,j}, 1 \le i \le n, 1 \le j \le m, 1 \le k \le 5$$

$$VARIABLES : t^{i,j} = \sum_{k=1}^{5} w_k * v_k^{i,j}, 1 \le i \le n, 1 \le j \le m$$
$$f^i, 1 \le i \le n$$

$$MINIMIZE : \sum_{i=1}^{n} f^i$$

$$SUBJECT\ TO : t^{i,j} - t^{i,1} < f^i * M, 1 \le i \le n, 2 \le j \le m$$
$$w_1 = 1$$
$$0 \le w_k, 2 \le k \le 5$$
$$0 \le f^i \le 1, 1 \le i \le n$$
$$f^i \text{ is an Integer}, 1 \le i \le n$$

Figure 1. MIP Model 1.

$$GIVEN : v_k^{i,j}, 1 \le i \le n, 1 \le j \le m, 1 \le k \le 5$$

$$VARIABLES : t^{i,j} = \sum_{k=1}^{5} w_k * v_k^{i,j}, 1 \le i \le n, 1 \le j \le m$$
$$f^{i,j}, 1 \le i \le n, 2 \le j \le m$$

$$MINIMIZE : \sum_{i=1}^{n} \sum_{j=2}^{m} f^{i,j}$$

$$SUBJECT\ TO : t^{i,j} - t^{i,1} < f^{i,j} * M, 1 \le i \le n, 2 \le j \le m$$
$$w_1 = 1$$
$$0 \le w_k, 2 \le k \le 5$$
$$0 \le f^{i,j} \le 1, 1 \le i \le n, 2 \le j \le m$$
$$f^{i,j} \text{ is an Integer}, 1 \le i \le n, 2 \le j \le m$$

Figure 2. MIP Model 2.

the extent of violation for a gene, i.e., the number of 'false' starts with better scores than the 'true' start for each gene. Intuitively, this should lead to a better discrimination in the test set. This slightly more sophisticated model is presented in Figure 2. The number of constraints in the second model grows by a factor of *m*.

The above models were implemented on Unix system using the AMPL modeling language (23) that interfaces with CPLEX as the underlying MIP solver. Training a system of equations with about 2000 constraints and four parameters to be trained, takes ~1 min on a Unix desktop DEC-alpha machine. The program is implemented in Perl and will be made available upon request along with any datasets used in this work.

RESULTS AND DISCUSSION

We first studied the distribution of distance between true start site and maximal start site. The nature of the distribution was very similar for all three species. The distribution for *B.subtilis* is shown in Table 2 as an example. Table 3 shows the distributions of start codons in the three genomes. Performance of our approach in start site prediction for two slightly different MIP models is summarized in Table 4.

Table 2. Distribution of distance between true start and maximal start for *B.subtilis*

Offset range (bp)	No. of genes
0–50	1017
51–100	151
101–150	54
151–200	10
201–250	6
250–300	3
>300	1

Table 3. Distribution of start codons in various genomes

Genome	*B.subtilis*	*E.coli*	*P.furiosus*
%ATG	79.9	94.0	90.8
%GTG	8.9	4.9	9.2
%TTG	11.2	1.1	0.0

To show the relative contribution of the factors used in the MIP models, we used each of these factors in isolation (excluding the distance between RBS and the start codon, since this does not make sense in isolation) to predict the start site. Performance of our approach is presented in terms of percentage true positives achieved in the training set and the complementary test set. When testing the factors in isolation, we use the entire 'true' gene set at our disposal since training the relative weights was not needed.

One of the two most prevalent tools for bacterial gene prediction, GLIMMER (3), picks the maximal ORF as the predicted gene. The other, GeneMark, lists the potential start sites for the user to choose from based on expert knowledge. One of our goals is to increase the reliability of the first-pass, computationally chosen start sites. The MAX_ORF line of Table 4 is then an indirect comparison to the use of the above tools in fully automatic mode, with no human post-processing.

To check the robustness of our approach, the system was trained on 10 different training sets containing a randomly chosen subset of the genes and tested on the remaining genes. The fraction of true positives (TP) on the training and the test set is represented by the mean and the standard deviation over the

Table 4. Performance comparison of various approaches

Genome		*B.subtilis*	*E.coli*	*P.furiosus*
No. of genes		1246	184	240
Fraction taken for training		0.2	0.5	0.5
MIP Model 1	TP% in training set (μ/σ)	92.6/1.9	93.6/1.5	92.7/1.7
	TP% in test set (μ/σ)	90.4/0.7	84.5/3.8	86.2/3.6
MIP Model 2	TP% in training set (μ/σ)	92.3/2.0	94.0/1.9	93.5/1.3
	TP% in test set (μ/σ)	90.4/0.8	84.9/4.0	86.6/3.2
MAX_ORF TP%		63	69	70
START_CODON TP%		67	81	82
RBS_BINDING_ENERGY TP%		85	59	64
START_PROBABILITY TP%		51	69	71

10 trials and is shown in Table 4. For each pair of training and test sets, the two models of MIP were applied (see Materials and Methods for the description of the models). Although the second model is more complex (with more constraints), it does not outperform the first model. We emphasize that we use approximately half of the genes as the training set in *E.coli* and *P.furiosus* and only ~20% of the genes in the case of *B.subtilis*. Traditionally, researchers have used at least 75% of the data set for training, which may exaggerate the performance of the method. From the results in Table 4, one might conclude that there is always one factor that is responsible for the combined performance but it is not uniform across species. For example, the RBS binding energy plays a critical role in *B.subtilis* start site prediction but not in the other two species. And the start codon alone is a reasonably accurate predictor of start site in *E.coli*. This remains the case if we disregard all but ATG as a valid start.

Each of the methods developed to date for improving start site prediction in prokaryotes has its own strengths, which we discuss next. It is difficult at this point to compare accuracy due to differences in testing methods. Any of the methods is probably much better than merely choosing the maximal ORF. The more ambitious user may want to use all three until the situation is clearer.

Obtaining a suitable set of reliably known gene starts for training and benchmarking is a difficult issue. Hayes (10) used the proteomics-derived *E.coli* set from (17), and the full *B.subtilis* annotations. There are ambiguities in interpreting the proteomics data due to post-translational processing, but on the whole they are quite reliable. The *B.subtilis* set of starts is much less certain since most of the annotations are computationally predicted with no experimental evidence. To avoid this circularity we only considered genes in *B.subtilis* with names starting with a 'non-y' character. These are the genes that have received individual attention. They may still have little supporting evidence for the particular initiation codon chosen, yet we feel it may be better than it seems since most investigators interested in a gene will have compared the gene carefully to known homologs. Training our algorithm on the 'non-y' genes and testing them on the 'y' genes gives an accuracy of ~84% implicating a lower accuracy of annotation in the 'y' genes. We used these two sets plus the *P.furiosus* set chosen by genome conservation with *P.horikoshii*. This last set is the most objective in that the reasoning is very direct—no deduced protein is involved—but also the least objective in that the genes to include were chosen on the basis of an 'obvious' change in conservation level. We feel it is actually very reliable, since the increase in conservation at the start site was, in fact, very clear in every case. This method of start prediction will be codified in an objective algorithm, tested carefully across a number of taxa, and published elsewhere. Frishman and colleagues start by searching PIR with criteria meant to eliminate most proteins predicted by computational means alone. These are then aligned to the genome and minimal ORFs around the aligned region are picked to compute the coding potential parameters. Then the ORFs (predicted *ab initio*) longer than certain threshold, with high coding potential and with only one start codon upstream of the aligned region are used as the training set for RBS profile generation. The training algorithm used has a unique strength in that it is resilient to inclusion of incorrect data, so that perhaps less care is needed in selecting a training set.

Each of the existing algorithms makes use of coding potential, but in different ways. Frishman and colleagues take the largest ORF with acceptable coding potential over the whole length, and then choose the 5'-most start with an acceptable RBS score. We, following Hayes (10), make use of the change in coding potential from before to after the initiation codon. All three approaches depend heavily on scoring a potential RBS. Hayes (10) and Frishman and colleagues (11) use a position weight matrix, while we use an estimate of binding energy. The matrix has the advantage of not requiring knowledge of the 16S rRNA, though in practice this is not an issue. The energy function has the advantage of being biologically based, not requiring any assumptions for training, and allowing bulges and loops in the secondary structure formed by the mRNA and the 16S rRNA. Though there are known cases where bulges and loops are almost certainly required (7), it is not known how important this consideration is in practice. To test the effect of loops and bulges, we applied the procedure on *E.coli* without allowing for loops and bulges. Notice that this changes both the binding energy and the distance of the binding site from the start codon. The average accuracy on the training and the test sets were 92.8 and 85.4% respectively using LP model 1. Repeating this on *B.subtilis* non-y genes resulted in

average accuracy of 91.93% on the training set and 89.7% on the test set. These results clearly de-emphasize the role of bulges and loops in binding sensitivity.

The majority of prokaryotic genes use the 5'-most start codon in the ORF. As noted above, Frishman and colleagues use the 5'-most start with a score above a minimum acceptable level. This approach divides all scores into only two classes. Our approach is to include the proximity to the 5' end of the maximal ORF as an explicitly scored feature of each potential start. Our algorithm explicitly scores potential start codons in terms of the frequency of occurrence of the codon itself. All three methods score the distance between the RBS and the start codon, but Frishman and colleagues build this into the RBS PWM, so that it cannot have an independent weighting factor. Hayes *et al*. (9) is the only method of the three to make use of the so-called 'downstream box' (24).

Hayes requires hand tuning in the optimization and, given the rate at which new genomes are appearing, we wanted an algorithm that could be optimized fully automatically. Frishman *et al*. choose a single universal cutoff for the combined score of the RBS and its distance from the initiation codon. On the other hand, we independently weight each of the five different features we use, and then choose these weights to optimize accuracy in the actual discrimination context, namely that of choosing one from among the first few potential starts in the maximal ORF.

The algorithm of Hayes (9) gives better accuracy on the *E.coli* test set (90.2%) and worse on the *B.subtilis* (80.6%). Dissecting the precise reasons for performance differences is extremely hard, since the two methods use independent sets of evidence. Difference in training/test set may well be part of it. The reason Hayes' approach does better with *E.coli* could be partially explained by: (i) he uses larger training set based on filtered set of genes from GeneMark: by using 80% of the known starts for training and testing on the remaining 20% the new approach achieves an average of 89.2%; and (ii) the performance figure shown is the average, the best performance achieved is 88.3%.

It is difficult to compare the accuracy of our method with that of Frishman *et al*. since the latter was evaluated against database annotation, which itself is often largely derived by computational methods. They report accuracy of 79.1% on the *B.subtilis* and 70.0% on the *E.coli* based on GenBank annotations.

In summary, we have presented a novel approach to bacterial start site prediction that takes into account multiple features of a potential start site. The methodology we have presented here is intended to supplement the existing bacterial gene finding tools by starting with what is probably a correct gene prediction and merely helping to pinpoint the particular initiation codon used. The algorithm includes a sufficiently comprehensive set of biologically motivated start site features that it should be applicable to a broad range of species. These features are combined using an elegant, appropriate and fully automatic optimization procedure.

ACKNOWLEDGEMENTS

We wish to thank V. Bafna and I. Ladunga for helpful discussion on the MIP modeling. We also wish to thank the reviewers for many insightful remarks that helped improve the paper greatly.

REFERENCES

1. Borodovosky,M. and McIninch,J. (1993) *Comp. Chem.*, **17**, 123–132.
2. Borodovosky,M., McIninch,J., Koonin,E., Rudd,K., Medigue,C. and Danchin,A. (1995) *Nucleic Acids Res.*, **23**, 3554–3562.
3. Salzberg,S., Delcher,A., Kasif,S. and White,O. (1998) *Nucleic Acids Res.*, **27**, 544–548.
4. Gelfand,M. (1995) *J. Comp. Biol.*, **2**, 87–115.
5. Stormo,G., Schneider,T. and Gold,L. (1982) *Nucleic Acids Res.*, **10**, 2997–3011.
6. Schurr,T., Nadir,E. and Margalit,H. (1993) *Nucleic Acids Res.*, **21**, 4019–4023.
7. Vellanoweth,R. (1993) In Losick,R. (ed.), *Bacillus subtilis and Other Gram-Positive Bacteria: Biochemistry, Physiology and Molecular Genetics*. American Society for Microbiology, Chapter 48.
8. Hatzigeorgiou,A. and Fickett,J. (1997) *Proc. First Annu. Conf. Comp. Genomics*, 8.
9. Hayes,W. and Borodovsky,M. (1998) *Proc. Pacific Symp. Biocomp.*, **3**, 279–290.
10. Hayes,W. (1998) Pattern recognition and signal detection in gene finding. Ph.D. thesis, Georgia Tech, School of Biology, Georgia Institute of Technology, Atlanta, GA.
11. Frishman,D., Mironov,A., Mewes,H.-W. and Gelfand,M. (1998) *Nucleic Acids Res.*, **26**, 2941–2947.
12. Fickett,J. (1996) *Trends Genet.*, **12**, 1058–1073.
13. Hayes,W. and Borodovsky,M. (1998) *Genome Res.*, **8**, 1154–1171.
14. Gelfand,M., Mironov,A. and Pevzner,P. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
15. Kunst,F., Ogasawara,N., Yoshikawa,H. and Danchin,A. (1997) *Nature*, **390**, 249–256.
16. Blattner,F., Plunkett III,G., Bloch,C., Perna,N., Burland,V., Riley,M., Collado-Vides,J., Glasner,J., Rode,C., Mayhew,G., Gregor,J., Davis,N., Kirkpatrick,H., Goeden,M., Rose,D., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1474.
17. Link,A., Robison,K. and Church,G. (1997) *Electrophoresis*, **18**, 1259–1313.
18. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) *J. Mol. Biol.*, **215**, 403–410.
19. Pearson,W. (1989) *Methods Enzymol.*, **183**, 63–98.
20. Freier,S., Kierzek,R., Jaeger,J., Sugimoto,N., Caruthers,M., Neilson,T. and Turner,D. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.
21. Beasley,J.E. (ed.) (1996) *Advances in Linear and Integer Programming*. Oxford Science Publication, Clarendon Press, Oxford.
22. Ladunga,I. (1999) *Bioinformatics*, **15**, in press.
23. Fourer,R., Gay,D. and Kernighan,B. (1993) *AMPL: A Modeling Language For Mathematical Programming*. Boyd and Fraser Publishing Company, Danvers, MA.
24. Sprengart,M.L., Fuchs,E. and Porter,A.G. (1996) *EMBO J.*, **15**, 665–674.