Nucleic Acids Research, 1998, Vol. 26, No. 12 **2941–2947**

Combining diverse evidence for gene recognition in completely sequenced bacterial genomes

Dmitrij Frishman*, Andrey Mironov[1], Hans-Werner Mewes and Mikhail Gelfand[2]

Munich Information Center for Protein Sequences (MIPS) of the German National Center for Health and Environment (GSF), Am Klopferspitz 18a, 82152 Martinsried, Germany, [1]Laboratory of Mathematical Methods, National Center for Biotechnology Information NIIGENETIKA, Moscow 113545, Russia and [2]Institute of Protein Research, Russian Academy of Sciences, Pushchino 142292, Russia

Recieved March 3, 1998; Revised and Accepted May 4, 1998

ABSTRACT

Analysis of a newly sequenced bacterial genome starts with identification of protein-coding genes. Functional assignment of proteins requires the exact knowledge of protein N-termini. We present a new program ORPHEUS that identifies candidate genes and accurately predicts gene starts. The analysis starts with a database similarity search and identification of reliable gene fragments. The latter are used to derive statistical characteristics of protein-coding regions and ribosome-binding sites and to predict the complete set of genes in the analyzed genome. In a test on *Bacillus subtilis* and *Escherichia coli* genomes, the program correctly identified 93.3% (resp. 96.3%) of experimentally annotated genes longer than 100 codons described in the PIR-International database, and for these genes 96.3% (83.9%) of starts were predicted exactly. Furthermore, 98.9% (99.1%) of genes longer than 100 codons annotated in GenBank were found, and 92.9% (75.7%) of predicted starts coincided with the feature table description. Finally, for the complete gene complements of *B.subtilis* and *E.coli*, including genes shorter than 100 codons, gene prediction accuracy was 88.9 and 87.1%, respectively, with 94.2 and 76.7% starts coinciding with the existing annotation.

INTRODUCTION

The principal goal of large-scale genome sequencing is to obtain new insights into physiological and biochemical processes in living organisms. An essential step in this process is gene identification with subsequent computer-based annotation of the corresponding gene products. Although bacterial genomic sequences are devoid of introns, gene recognition in bacteria is far from being simple. It is easy to extract all possible open reading frames (ORFs) from a given DNA sequence; it is much less trivial to decide which of them correspond to genes that are actually expressed and code for proteins. The following features are important indicators of protein coding regions in DNA: (i) sufficient ORF length. Long ORFs rarely occur by chance; (ii) specific patterns of codon usage that are different from triplet frequencies in non-coding regions ('coding potential'); (iii) the presence of ribosome binding sites (RBS) in the (–20)...(–1) region upstream of the start codon that help to direct ribosomes to the correct translation start positions (1). A part of the RBS is formed by the purine-rich Shine–Dalgarno (SD) sequence which is complementary to the 3′ end of the 16S rRNA (2); (iv) similarity to known, especially experimentally characterized, gene products.

Correspondingly, the approaches to gene recognition are traditionally divided into two broad categories (3,4). Intrinsic, or *ab initio* methods utilize statistic, linguistic or pattern recognition algorithms to find genes in DNA through detection of specific nucleic acid motifs or global statistical patterns, whereas extrinsic methods take into account information about other known proteins.

There exist numerous algorithms for *ab initio* recognition of protein-coding regions and functional sites (reviewed in 5,6). The most popular gene prediction program for prokaryotes, GeneMark (6), utilizes non-homogeneous Markov models to find DNA regions that code for proteins or are complementary to them. Non-coding regions are described by homogeneous Markov models. A Bayesian decision rule is applied to deduce the coding capacity of sliding windows. GeneMark has been used in several genome sequencing projects (e.g. 7–9). Recently this algorithm was extended to take into account information about candidate ribosomal binding sites (10). The recently developed GLIMMER (11) has been reported to provide very high gene prediction accuracy in *Haemophilus influenzae* and *Helicobacter pylori* genomes. GLIMMER relies on interpolated Markov models to take into account DNA oligomers of varying length, dependent on the local composition of the sequence. Another program, EcoParse (12), utilizes hidden Markov models to find the maximum likelihood parse of a DNA sequence into coding and non-coding regions without the use of sliding windows. A program for gene recognition relying solely on ORF length and RBS was described by Hatzigeorgiou and Fickett (13).

Extrinsic analysis involves similarity searches with candidate gene products against protein sequence databanks. The most popular program of this class, BLASTX (14), performs six-frame translation of the query DNA and compares the resulting amino

*To whom correspondence should be addressed. Tel: +49 89 8578 2664; Fax: +49 89 8578 2655; Email: frishman@mips.biochem.mpg.de

acid sequence to known proteins. Search results are represented in an integrated report, with hits from different reading frames combined to produce one statistically meaningful similarity score. Robinson *et al.* (15) used BLASTX to detect 450 new bacterial genes missed in original publications, including several genes previously known only in eukaryotes. Another DNA–protein search program, DPS (16), is the only currently available software tool that allows us to compare a complete genome sequence (3–5 mbp and more) with the total protein sequence databank in one pass.

Neither extrinsic nor intrinsic methods taken separately can ensure successful prediction. Practical experience in prokaryotic genome analysis as well as the recent trends in gene recognition in higher organisms (17,18) show that it is necessary to incorporate all available evidence in order to achieve reliable results. In real life, putative coding regions predicted by intrinsic methods are verified by similarity searches. Finding a related protein serves as a decisive supporting evidence. Pearson *et al.* (19) studied the ORFs predicted with GeneMark in *H.influenzae*, *Methanococcus jannaschii* and *Mycoplasma genitalium*. In many cases they were able to correct the length of genes based on comparative analysis with known proteins. Additionally they found many short genes not identified by GeneMark. The overall conclusion of this work is that a sizeable amount of genes annotated within the framework of large-scale sequencing projects are fully or partially wrong.

Experience coming from many computational genome analysis efforts (20–22) shows that 60–80% of genes in newly sequenced organisms have known counterparts in other species. In many cases the similarity is only marginal, partial or to gene products without known function. However, in at least 30% of the cases, reliable global alignments with well characterized proteins can be obtained. We were thus tempted to invert the usual procedure in which genes predicted by *ab initio* statistical methods are accepted or rejected based on subsequent similarity searches. In this work we use DNA regions significantly related to known proteins to extract codon usage statistics and other intrinsic recognition parameters that are further applied to unexplored parts of a genome. The leading idea of this work is that extrinsic evidence should be given higher priority than intrinsic information.

We also pay specific attention to assignment of gene starts. This is important since 5′ ends of genes often are not conserved, whereas they carry important functional and structural information. In particular, a signal peptide may provide information about protein localization (23). The N-terminus can contain information about the life span of a protein (24). The estimated strength of the RBS can be an indicator of the efficacy of the translation initiation (25). Thus correct determination of the gene start can be as important as identification of the gene itself.

MATERIALS AND METHODS

Data

Complete nucleotide sequences of the *Bacillus subtilis* (9) and *Escherichia coli* (8) genomes were downloaded from the SubtiList WWW Server at the Pasteur Institute (http://www.pasteur.fr/Bio/SubtiList.html) and the *E.coli* genome project resource at the University of Wisconsin (http://ecoliftp.genetics.wisc.edu/), respectively. In addition, we also obtained full sets of the protein sequences encoded in these two genomes (4099 for *B.subtilis* and 4277 for *E.coli*) as assigned by the genome authors based on the application of various computational techniques as well as manual analysis.

Independently, *B.subtilis* and *E.coli* sequences were extracted from the PIR-International protein sequence database using the Sequence Retrieval System (SRS; 26). Special care was taken to select only sequences determined by individual researchers in different laboratories not associated with large-scale sequencing projects. Sequences submitted after 1995, plasmid sequences, fragments and proteins described in PIR as hypothetical, as well as the PIR entries containing the names of the main researchers involved in the *B.subtilis* and *E.coli* genome projects were discarded. These two sets were compared with the full sets of gene products from the two genomes using the BLAST2 software (W.Gish, unpublished; 27,14). Only the PIR sequences at least 98% identical to their counterparts in complete genomes and having the same N-terminal sequence were retained. This selection procedure resulted in 219 *E.coli* and 346 *B.subtilis* proteins.

Throughout this work, the full sets of gene products from complete genomes as determined by the authors will be referred to as SUBGEN (for *B.subtilis*) and ECOGEN (for *E.coli*), and the sequence sets extracted from the PIR database as SUBPIR and ECOPIR, respectively.

For similarity searches we created a non-redundant protein databank by merging the PIR-International (28), SWISS-PROT, SWISS-NEW, TREMBL and TREMBLNEW (29) sequence collections using the NRDB2 software developed by W.Gish (unpublished). Sequences from all species for which genome sequencing projects have been completed were excluded. The resulting databank currently contains 208 660 protein sequences.

Outline of the algorithm

Our algorithm is based on the assumption that information about coding regions derived from similarity searches is in principle more reliable than statistical data. We use the term ‘seed ORF’ to describe the minimal, most reliable possible ORF that can be inferred. In the case of similarity searches, a seed ORF is obtained by extending the reliably aligned region in the upstream direction until the first start codon occurs and in the downstream direction until a stop codon is encountered. These similarity-derived seed ORFs are used to calculate coding potential parameters. For ORFs predicted *ab initio* a seed ORF results from extending a DNA region of a given minimal length (e.g., 300 nt) possessing sufficiently high coding potential (see below) in the same fashion. At the next step of analysis the algorithm tries to extend the seed ORFs by including additional upstream DNA fragments encompassing the next available start codon provided that the DNA region between the old and new candidate starts satisfies conditions imposed on coding potential. The sample of ORFs with a single possible start codon is used to derive the RBS recognition matrix. Finally, in ORFs with multiple candidate starts, the leftmost start codon having sufficiently strong RBS is selected.

Similarity search and the set of seed ORFs

We used the DPS program (16) to compare complete genomic sequences with the complete non-redundant protein sequence databank. DPS performs mapping of all protein sequences from the database onto the query genomic sequence. The DPS output contains full information about a DNA–protein match, including the start and end positions, reading frame, similarity score and

alignment of the high-scoring DNA segment with the corresponding protein sequence fragment represented in three-letter code. The alignment may be split into several high-scoring fragments, in which case reading frames, coordinates and similarity scores are given for each such fragment, and the aggregate similarity for the entire alignment is indicated separately. We took into account only DPS hits with sufficiently high aggregate scores (typically >750) involving only one reading frame; cases involving more than one reading frame are subject to a separate procedure aimed at detecting frame-shifts.

Coding potential and the complete set of candidate ORFs

Seed ORF sets produced by the similarity search were utilized to calculate the codon usage tables and the average and standard deviation of the coding potential. To do that, the significant DNA–protein alignments were extended until the first stop codon downstream and the first start codon upstream occurred. The obtained DNA sequences are the most reliable representatives of the coding parts of the genome that can be extracted automatically.

Let $F(abc)$ be the genomic frequency of the codon abc. Statistical weight of abc is defined as $W(abc) = \log F(abc)$. *Primary coding potential* of a DNA segment of length n codons is its log-likelihood:

$$Q(a_1b_1c_1...a_nb_nc_n) = \sum_{k=1}^{n} W(a_kb_kc_k).$$

To account for DNA fragments of different length, we will use the normalized potential measured in the standard deviation units:

$$R = \frac{Q-\mu n}{\sigma\sqrt{n}},$$

where μ is the average codon weight which depends on the base composition of the genome and σ is the standard deviation:

$$\mu = \sum_{b_1b_2b_3=\mathrm{AAA}}^{\mathrm{TTT}} G(b_1)G(b_2)G(b_3)W(b_1b_2b_3),$$

$$\sigma^2 = \sum_{b_1b_2b_3=\mathrm{AAA}}^{\mathrm{TTT}} G(b_1)G(b_2)G(b_3)[W(b_1b_2b_3) - \mu]^2.$$

[$G(b)$ is the genomic frequency of the base b].

Finally, to avoid the influence of the local base composition and gene shadows, and to set the strand and reading frame, we define the *coding quality* of a DNA fragment as

$$\Omega(a_1b_1c_1...a_nb_nc_n) = R(a_1b_1c_1...a_nb_nc_n) - \max\{R(c_0a_1b_1...c_{n-1}a_nb_n), R(b_1c_1a_2...b_nc_na_{n+1})\}.$$

Upon derivation of the statistical parameters above, the DPS output was screened again to extract all similarity-based seed ORFs. The parts of the genome not covered by the similarity-based seed ORFs were subsequently analyzed for the presence of other protein-coding seed ORFs. A seed ORF was accepted if its length exceeded a given threshold (100 codons) and its coding quality Ω was sufficiently high.

All seed ORFs were then extended in the 5′ direction as far as possible. Short overlaps between genes (up to 6 nt) were allowed. Each extension piece of DNA started with ATG, GTG or TTG. The extension was accepted if the coding quality Ω of the DNA segment of length 99 nt starting with the new candidate start codon was acceptable; otherwise the extension of a given ORF was interrupted. The window length of 99 was chosen to ensure sufficient statistical significance of the calculated coding potential (30).

This procedure resulted in the complete set of ‘open-start’ candidate ORFs. The start codons for this set were assigned at the final step.

RBS weight matrix and assignment of gene starts

Candidate ORFs with only one possible start codon and not having neighbors closer than 30 bases upstream were selected. Regions (–20)...(+3) of these ORFs were aligned at start codons. These sequences were used to derive the RBS weight matrix.

Let L be the expected length of the SD box ($L = 6$). Denote by $F(b,j)$ positional nucleotide frequencies in the initial alignment [$j = (-20)...(-1)$; b = T,C,A,G]. Positional information content is:

$$H(j) = \sum_{b=\mathrm{A}}^{\mathrm{T}} F(b,j)\log(F(b,j)/G(b)),$$

$G(b)$ is the genomic frequency of the base b (31).

Initially the RBS signal was assumed to reside in positions having the maximum total information content:

$$\sum_{k=j}^{j+L-1} H(k) \xrightarrow[j\,=\,-20...-L]{} max.$$

Then the position of the SD box in each individual sequence was determined using the following two-stage procedure.

We start with some definitions. Denote by $N(b,k)$ positional nucleotide counts in the SD profile at a given iteration ($k = 1...L$). Positional nucleotide weights are:

$$W(b,k) = \log\left(\frac{N(b,k) + 0.5}{\max_b N(b,k) + 0.5}\right),$$

$\max N(b,k)$ is the frequency of the consensus nucleotide in the position k. The SD signal score is calculated by the formula:

$$\Delta(b_1...b_L) = \sum_{k=1}^{L} W(b_k,k).$$

The first re-alignment stage involves the iteration until convergence of the following two steps: (i) find in each sequence the segment of length L with the highest score; (ii) re-calculate the nucleotide weight matrix.

A distinctive feature of our algorithm is that at each optimization step only the top scoring fraction (usually top 80%) of sequences are used to produce the current weight matrix.

At the second stage the preferences for the distance between the SD box and the start codon are taken into account. Let M be a possible position of the SD box within the RBS region, and let this position occur $N(M)$ times. Denote by $N_{\max}$ the count of the most frequent position. The positional weights are calculated using the standard formula:

$$V(M) = \log\left(\frac{N(M) + 0.5}{N_{\max} + 0.5}\right).$$

Now the strength of the SD signal is defined as:

$$\Delta(b_1...b_L) = \sum_{k=1}^{L} W(b_k, k) + V(M).$$

Thus the RBS profile is the nucleotide weight matrix and the vector of position weights. The two step iterative procedure is used again until convergence.

The final step of the genome annotation is the assignment of start codons to 'open start' candidate ORFs. If a candidate ORF contains start codons with sufficiently strong RBS, the 5′-proximal of these starts was accepted. Otherwise the initial start generated at the previous stage was used, thus taking into account the possibility of translation re-initialization from an upstream gene.

Choice of the minimal allowed ORF length and handling of short ORFs

First versions of our program worked with a fixed minimal ORF length, typically 100 codons. The conflicts between overlapping ORFs were resolved based on the strength of the coding potential, as described above. The disadvantage of this approach was that quite often short ORFs defeated much longer competing ORFs, and the genome regions vacated by the latter would be returned to the pool of unoccupied space, giving rise to additional abundant short ORFs. This problem became especially severe when the minimal allowed ORF length was set to values under 100 codons, which led to a large number of predicted short ORFs at the expense of the longer ones.

To resolve this problem, we modified the final stage of gene prediction process as follows. The minimal allowed ORF length is first set to a very high value (2000 nt), after which all genes are predicted as described above. Then the minimal ORF length is reduced step-wise, and the gene prediction process repeated. At that, the genes predicted at the previous step remain unchanged. Thus, longer ORFs get higher priority, and the next pool of ORFs is derived from the genome regions that are unoccupied after completion of the previous step. This allows us to avoid the explosion of short ORFs while preserving high overall prediction accuracy.

Table 1. RBS weight matrix for *B.subtilis* and *E.coli*

Nucleotide	Nucleotide position in the window					
	1	2	3	4	5	6
Bacillus subtilis						
A	0.000	–0.909	–0.845	0.000	–0.909	–0.511
C	–0.856	–1.000	–0.943	–0.923	–0.999	–0.748
G	–0.804	0.000	0.000	–0.709	0.000	0.000
T	–0.765	–0.897	–0.980	–0.868	–0.962	0.725
Consensus	A	G	G	A	G	G
Escherichia coli						
A	0.000	–0.035	0.000	–0.995	–0.936	0.000
C	–0.299	0.000	–0.804	–0.984	–0.987	–0.814
G	–0.386	–0.115	–0.903	0.000	0.000	–0.710
T	–0.027	–0.426	–0.752	–0.937	–1.009	-0.749
Consensus	A/T	C/A	A	G	G	A

Implementation and availability

The program to calculate weight matrices based on a multiple alignment of putative RBS regions is called STARTER and is written in C programming language. All other computational steps described in this paper are implemented as a Perl 5 script called ORPHEUS. Both programs as well as all data mentioned (protein sequence sets, weight matrices, DPS search results, etc.) are freely available to academic users; see http://pedant.mips.biochem.mpg.de/frishman/orpheus_home.html

RESULTS AND DISCUSSION

910 *E.coli* and 529 *B.subtilis* similarity-based seed ORFs were extracted from the DPS search results. The average length of these ORFs was 444 and 507 amino acids, respectively, greater then the average length of *E.coli* (339) and *B.subtilis* (326) genes (excluding the genes shorter than 100 amino acids). The reason for this difference is that a very stringent DPS similarity score threshold was chosen, giving preference to long, reliable alignments.

RBS weight matrices (Table 1) were derived from alignments of 385 *B.subtilis* and 644 *E.coli* 5′ upstream gene regions with single candidate starts (Fig. 1). The optimization algorithm converged after 30–40 iterations. As seen in Figure 2, certain locations of the candidate SD box relative to the start codon, in this case –13 both in *B.subtilis* and *E.coli*, are strongly preferred. Thus incorporation of the positional preference information in

```
1    AGAAAACGACAAGGAAAGGTATTTG   -1.763
2    ATCTGAAGGGGGATTTTGGAGAATG   -0.965
3    GTGAAAAATTGGAGGGAAACTCATG   -0.765
4    CAATTAGAGAGGAGAATTCGATATG   -0.511
5    AAAGCGAAGGACTCGGCGAGTAATG   -1.884
6    ACATACCCTGCAAGGATGATTAATG   -1.264
7    AAAAGGAAGGGAGGTCTATCTCATG   -0.914
8    TTAAATATGGTGGTGGAAACAGATG   -1.793
9    TTTCGAAAGGATTGTTTATAAAATG   -1.847
10   GACTGAATATTTAAACAATTATGTG   ------
11   TAAATAGAAGGAGGCGCACAAAATG   -0.110

...

376  ACCGTTATGGAGGGATACATAAATG   -0.925
377  TAACTGAATTTAAAGGAGGTTCATG   -0.325
378  CGTTTAAAATGCATAATAAGGAGTG   -2.061
379  TGGAAAACAGGGAGAGATCATAATG   -1.315
380  TGCAAACTAACGGGGGGATAATATG   -1.710
381  AAAAGAAAAAGGAGATGGGAGTATG   -0.511
382  ATACAAGGTCTTTCGGGAGGCCTTG   -1.159
383  ATCATGATCGGTCATATTTTAGATG   ------
384  CGAATGTAAACATGTAGCAAGGGTG   -2.171
385  AAATTCGGGAGAGTGAAGCGAGATG   -1.570
386  ACAAACAGAATTCAGGTGAGACATG   -1.704
```

Figure 1. Alignment of the *B.subtilis* regions upstream from the 5′ ends of the ORFs with one possible start codon and acceptable coding potential. Sequences are numbered 1–386. Each sequence includes positions –22...–1 upstream of the start codon and the start codon (shown in italic). Location of the regions with the highest RBS score are shown by boxes, and the corresponding RBS scores are indicated in the last column. The location of the preferred SD position (in this case –13; see Fig. 2) is shaded. Note that the procedure to find RBS uses the top scoring 80% of sequences at each iteration step. Sequences with the worst 20% of scores (in this example 10 and 383) are ignored.

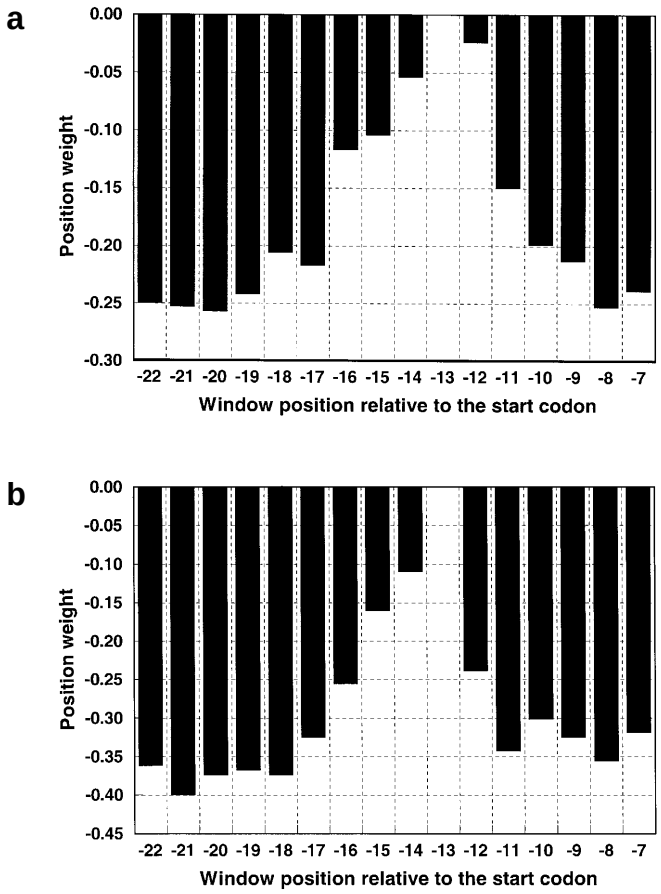

Figure 2. Automatically derived positional preference of the SD box in *B.subtilis* (**a**) and *E.coli* (**b**). Higher values (closer to 0) correspond to more preferred start positions of the window of length 6 nt.

Start codon position in the genome	RBS signal strength	Coding potential quality	Putative translation initiation region
3252280	-4.171	2.090	gaatcctcattgtaaaattatcGTG
3252331	-3.423	1.031	tctaggcggcgaggtcaatgggATG
3252364	-3.684	0.927	ctcgtcatatgatctcattttaATG
3252445	-2.388	0.484	aattttggaaacggattcgaatTTG
3252463	-3.738	0.312	catggaaggcaccaagacaatttTG
3252484	-3.950	0.777	gattgatgaccatccggctgtcATG
3252508	-2.706	0.780	ggaaaacatgaaaaagatactaGTG
3252523	**-0.718**	**0.844**	**agtgagtaaaagggaggaaaacATG**
3252544	-2.502	0.698	ctttttataaaatggaaaagaGTG

Figure 3. Start codon selection procedure. The seed ORF of the *comA* gene (32) with multiple possible starts situated on the complementary strand is extended in the upstream direction, and the values of the coding potential downstream of each start codon (Ω) and RBS signal strength upstream (Δ) recorded. Start position 3 252 523 is selected since it is preceded by a very strong RBS (bold line). The SD sequence indicated in (32) is underlined. Start codons are shown in upper case. Note that the values of Ω are higher than the conservative threshold –1.0 in all cases.

addition to the standard weight matrices allows for higher selectivity in RBS detection.

The start selection procedure is exemplified in Figure 3 using the sequence of the *B.subtilis* competence gene *comA* (PIR code RGBSCA) originally determined by Weinrauch *et al.* (32). This protein has no strong similarity to proteins of other species and thus its seed ORF was not similarity based. Figure 3 illustrates the search for an appropriate start position of the *comA* gene after its seed ORF starting in position 3 252 280 of the genome (on the complementary strand) was detected. The seed ORF was then extended in the 5′ direction (increasing position numbers), and the values of the coding quality Ω and RBS strength Δ were recorded. As seen in Figure 3, the ATG start codon in position 3 252 523 was accepted since it has a very strong RBS upstream (Δ = –0.718, much higher than the recognition threshold –2.0 used for *B.subtilis*). This corresponds to the gene start position identified by Weinrauch *et al.* (32).

Our program identified 4379 genes longer than 35 codons in the *B.subtilis* genome and 4595 genes longer than 35 codons in the *E.coli* genome. As seen in Table 2 and Figure 4, most of the false negatives, i.e. genome proteins not identified by our program, and false positives, i.e. over-predicted ORFs, are shorter or slightly longer than 100 codons. The numbers of predicted ORFs longer than 100 codons were 3555 in *B.subtilis* and 3724 in *E.coli*, very close to 3613 genes in *B.subtilis* and 3901 in *E.coli* genome determined in the original publications. As seen in Table 2, <2% of predicted genes longer than 100 codons had no counterparts in published genome data. The agreement with the SUBPIR and ECOPIR subsets was also quite good, with <4% false positives. Over 90% of predicted *B.subtilis* genes had correct start positions, with most of the length differences for other genes under 10 codons (data not shown). In *E.coli* gene start positions were predicted with much smaller confidence (>75%). For comparison we also present results of the widely used ‘leftmost ATG’ procedure for the start codon assignment (Table 2). It appears that our start detection procedure is not efficient for *E.coli* and probably for other genomes with weak RBS signals, while in *B.subtilis* our algorithm substantially outperformed the ‘leftmost ATG’ rule.

As seen in Figure 4, our algorithm is capable of handling short ORFs in the range 80–100 codons with reasonable accuracy. For shorter ORFs the prediction quality quickly deteriorates, making automatic detection of very short peptides nearly impossible. However, the main source of errors in the length range between 20 and 80 codons is false positives, or in other words, unsupported ORFs for which no experimental evidence proving their existence is available. We can not exclude that some, or even many of these putative genes may be real.

The main distinctive feature of our algorithm is that we start the analysis with the coding regions and candidate RBS that can be expected to be highly reliable. They serve as a learning set used to derive statistical parameters used for further, more detailed analysis. The use of top 80% highest scoring candidate RBS to derive the weight matrix proved to be highly effective and improved discrimination power of the weight matrix.

Unlike GeneMark and EcoParse, our algorithm does not rely on the statistics of the non-coding regions. This is motivated by the fact that only coding regions can be defined unambiguously, especially at the initial steps of the analysis. Similarly, we do not use the energy of the base-pairing of the SD and the 16s rRNA. This makes the program applicable at early stages of genome analysis when the rRNA genes may be not be sequenced yet. Also, in some bacteria the RBS does not conform to the standard base-pairing model (33). Finally, we do not use complicated multiple alignment techniques for derivation of the RBS profile: it turned out that the relatively strong RBS signal can be detected by a relatively simple iterative procedure (cf. 34).

Table 2. Comparison of the gene prediction results with the sets of sequences from the PIR-International and the genome sequencing projects

Dataset	% correctly identified genes (true positives)		% correct starts for correctly identified genes		% correctly predicted genes with correct starts using 'leftmost ATG' procedure	
	L > 100	L > 35	L > 100	L > 35	L > 100	L > 35
SUBPIR	93.3	–	96.3	–	83.0	–
ECOPIR	96.3	–	83.9	–	86.9	–
SUBGEN	98.9	88.9	92.9	94.2	75.7	82.8
ECOGEN	99.1	87.1	75.7	76.7	78.0	77.7

L, length.

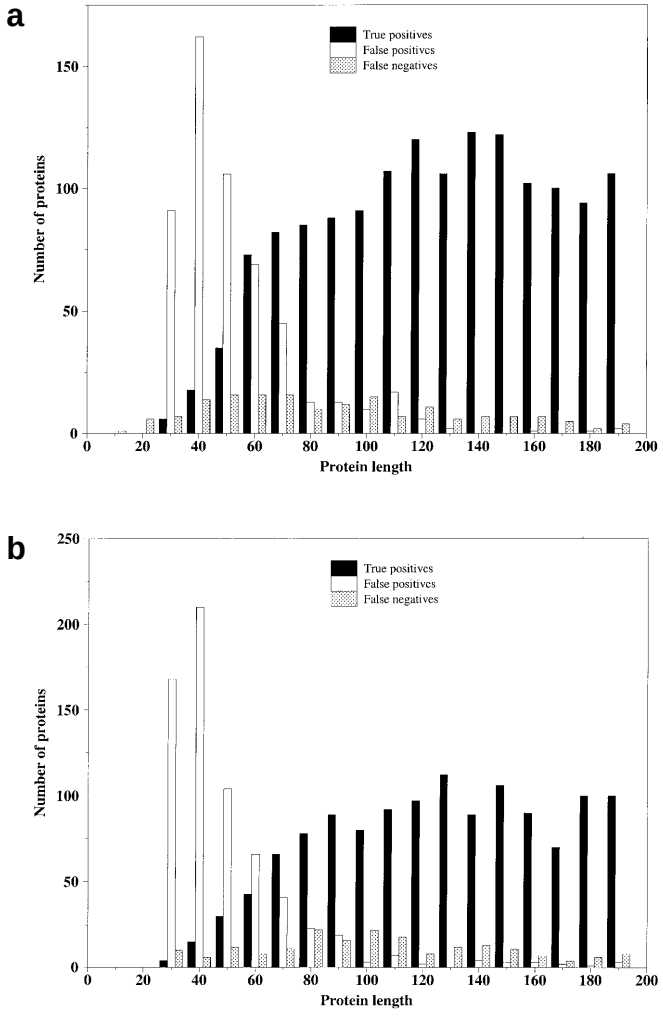

Figure 4. Distribution of the true positive, false positive and false negative lengths in *B.subtilis* (**a**) and *E.coli* (**b**) ORFs. Only the ORF length range 0–200 codons is shown. Predictions for longer ORFs are practically perfect.

Most genes in the current databanks, and specifically the genes determined in the framework of major sequencing projects, are not corroborated experimentally. This makes it very difficult to assess performance of any particular algorithm or perform large-scale benchmarking (cf. the detailed discussion in 13). Interestingly, the gene start prediction accuracy both in *B.subtilis* and *E.coli* was a few percentage points higher for the SUBPIR and ECOPIR subsets than for the full genomes, whereas for the percentage of correctly predicted genes the situation was the opposite. The differences between the prediction results for the PIR and GenBank data sets can be explained by the details of the gene analysis in the original publications. However, analysis of the disagreements between different annotations should be undertaken in order to resolve this problem.

A slightly worse percentage of identified genes in *B.subtilis* as compared to *E.coli* can be explained by the relatively uniform codon usage in these species (35; see, however, 36). On the other hand, much better assignment of gene starts in *B.subtilis* reflects the general tendency towards stronger RBS in some Gram-positive bacteria (13).

The second surprising finding is the large number of non-ATG start codons. Indeed, in the candidate genes of *B.subtilis* having only one possible start codon, this codon is TTG in 21% of genes and GTG in 16% of genes (cf. 13 and 9% respectively, in 9). Since it is unlikely that the set of similarity-seed ORFs is somehow biased in the use of start codons, we feel that the former values are likely to be correct.

The future direction of this work is to incorporate additional evidence for better prediction of genes and their starting positions. It would be very desirable to take into account the influence of the mRNA secondary structure on the choice of start codons (e.g. 37) and to mask the genome regions coding for stable RNAs such as rRNAs and tRNAs (38) in order to decrease the number of false positives. Protein features can also be important for gene recognition. At present the gene recognition programs serve mainly as an initial step of genome analysis, to be followed by protein functional and structural analysis (22). An obvious possibility is the use of signal peptide predictions for the choice of the start codon. However, more sophisticated uses of protein analysis are possible. This probably should be done by hierarchical analysis systems with various feedback connections.

ACKNOWLEDGEMENTS

We are grateful to M.Galperin, A.Grigoriev, J.Hani, E.Koonin, P.Pevzner and M.Roytberg for useful discussions and to A.Hatzigeorgiou and J.W.Fickett for communicating their results prior to publication. A.M. and M.G. are partially supported by grants from the Russian Fund of Basic Research, the Russian State Program 'Human Genome' and from the USA Department of Energy (DE-FG-94ER61919). The support of the

Bundes-ministerium für Forschung und Technologie (FKZ 0311670) is gratefully acknowledged.

REFERENCES

1 Dreifus, M. (1988) *J. Mol. Biol.*, **204**, 79–94.
2 Shine, J. and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA*, **71**, 1342–1346.
3 Borodovsky, M.Y., Rudd, K.E. and Koonin, E.V. (1994) *Nucleic Acids Res.*, **22**, 4756–4767.
4 Fickett, J.W. (1996) *Trends Genet.*, **12**, 316–320.
5 Gelfand, M.S. (1995) *J. Comput. Biol.*, **2**, 87–115.
6 Borodovsky, M.Y. and McIninch, J.D. (1993) *Comput. Chem.*, **17**, 123–133.
7 Bult, C.J., *et al.* (1996) *Science*, **273**, 1058–1073.
8 Blattner, F.R., Plunkett, G., III, Bloch, C.A., *et al.* (1997) *Science*, **277**, 1453–1462.
9 Kunst, F., *et al.* (1997) *Nature*, **390**, 249–256.
10 Lukashin, A.V. and Borodovsky, M. (1998) *Nucleic Acids Res.*, **26**, 1107–1115.
11 Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) *Nucleic Acids Res.*, **26**, 544–548.
12 Krogh, A., Mian, I.S. and Haussler, D. (1994) *Nucleic Acids Res.*, **22**, 4768–4778.
13 Hatzigeorgiou, A.G. and Fickett, J.W. (1997) Locating translation initiating sites in *Bacillus subtilis* and other Gram-positive bacteria. *First Annual Conference on Computational Genomics*, p.8.
14 Gish, W. and States, D.J. (1993) *Nature Genet.*, **3**, 266–272.
15 Robinson, K., Gilbert, W. and Church, G.M. (1994) *Nature Genet.*, **7**, 205–214.
16 Huang, X. (1996) *Microb. Compar. Genomics*, **1**, 281–291.
17 Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 9061–9066.
18 Burge, C. and Karlin, S. (1997) *J. Mol. Biol.*, **268**, 78–94.
19 Pearson, W., Wood, T., Zhang, Z. and Miller, W. (1997) *Genomics*, **46**, 24–36.
20 Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E.L.L. (1992) *Nature*, **358**, 287.
21 Koonin, E.V., Muschegian, A.R. and Rudd, K.E. (1996) *Curr. Biol.*, **6**, 404–416.
22 Frishman, D. and Mewes, H.W. (1997) *Trends Genet.*, **13**, 415–416.
23 Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) *Protein Engng.*, **10**, 1–6.
24 Varshavsky, A. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 12142–12149.
25 Barrick, D., Villanueba, K., Childs, J., Kalil, R. and Schneider, T.D. (1994) *Nucleic Acids Res.*, **22**, 1287–1295.
26 Etzold, T., Ulyanov, A. and Argos, P. (1996) *Methods Enzymol.*, **266**, 114–128.
27 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
28 Barker, W.C., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S.L., Ledley, R.S., Mewes, H.W., Pfeiffer, F. and Tsugita, A. (1998) *Nucleic Acids Res.*, **26**, 27–32.
29 Bairoch, A. and Apweiler, R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
30 Fickett, J.W. and Tung, C.S. (1992) *Nucleic Acids Res.*, **20**, 6441–6450.
31 Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.*, **188**, 415–431.
32 Weinrauch, Y., Guillen, N. and Dubnau, D.A. (1989) *J. Bacteriol.*, **171**, 5362–5375.
33 Hayes, W.S. and Borodovsky, M. (1998) In Altman, R.B., Dunker, A.K., Hunter, L. and Klein, T. (eds), *Pacific Symposium on Bioinformatics '98*, pp. 279–290.
34 Golovanov, E.I., Sprizhitsky, Yu.A. and Alexandrov, A.A. (1982) *Abstr. 6th Symp. 'Structure and Functions of Proteins and Nucleic Acids'*, p 52, Tskhaltubo.
35 Ogasawara, N. (1985) *Gene*, **40**, 145–150.
36 Shields, D.C. and Sharp, P.M. (1987) *Nucleic Acids Res.*, **15**, 8023–8040.
37 Kister, A.E. (1990) In Frank-Kamenetsky, M.D. (ed.), *Computer Analysis of Genetics Texts*, Moscow. pp 189–220.
38 Eddy, S.R. and Durbin, R. (1994) *Nucleic Acids Res.*, **22**, 2079–2088.