Gene recognition in completely sequenced bacterial genomes [1]

I Introduction

Multiple bacterial sequencing projects created the new computational challenge of *in silico* * gene prediction in the absence of any experimental analysis. The problem is that in the absence of experimentally verified genes, there are no positive or negative test samples from which to learn the statistical parameters for coding and non-coding regions. Frishman et al., [1] proposed the "similarity-first" approach, which first finds fragments in bacterial DNA that are closely related to fragments from a database and uses them as the initial training set for the algorithm. After the statistical parameters for genes that have related sequences are found, they are used for prediction of other genes in an iterative fashion.

Although bacterial genomic sequences are devoid of introns**, gene recognition in bacteria is far from being simple. It's easy to extract all possible open reading frames*** (ORFs) from a given DNA sequence, but it is much less trivial to decide which of them correspond to genes that are actually expressed and code for proteins.

Features as important indicators of protein coding regions in DNA are:

(i) sufficient ORF length. Long ORFs are 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) regions upstream of the start codon that help to direct ribosome to the correct translation start positions

(iv) similarity to known, especially experimentally characterized, gene products.

Approaches to gene recognition are traditionally divided into intrinsic and extrinsic approaches. 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 (programs as GeneMark, GLIMMER, EcoParse), whereas extrinsic methods involves similarity searches with candidate gene products against protein sequence databanks (programs as BLASTX, DPS). Actually, neither extrinsic nor intrinsic methods taken separately can ensure successful prediction, thus it is necessary to incorporate all available evidence in order to achieve reliable results. In this work, authors use DNA regions significantly related to known proteins to extract codon usage statistics and other intrinsic recognition parameters that are further applied to

* The use of computers to simulate, process, or analyse a biological experiment.

** Nucleotide sequences found in the structural genes of eukaryotes that are non-coding and interrupt the sequences containing information that codes for polypeptide chains. Intron sequences are spliced out of their RNA transcripts before maturation and protein synthesis.

*** Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon, and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.

unexplored parts of a genome. The leading idea of this work is that extrinsic evidence should be given higher priority that intrinsic information.

II Methods

[Data] Nucleotide sequences of the *Bacillus subtilis* and *Escherichia coli* were extracted from the PIR-International protein sequence database using the Sequence Retrieval System [2] with some special care. These two sets were compared with the full sets of gene products from the two genomes using the BLAST2 [3]. Only the PIR sequences at least 98% identical to their counterparts in complete genomes and having the same N-terminal sequences were retained. This selection procedure resulted in 346 *B. subtilis* and 219 *E. coli* proteins.

For similarity searches the authors created a non-redundant protein databank by merging the PIR-International, SWISS-PROT, SWISS-NEW, TREMBL and TREMBLNEW sequence collections using NRDB2 software (by W. Gish).

[Algorithm] The assumption of this algorithm is that information about coding regions derived from similarity searches is in principle more reliable than statistical data.

*** Similarity search and the set of seed ORFs** The authors used the DPS [4] 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 sequences. This work took into account only DPS hits with sufficiently high aggregate scores involving only one reading frame. 'Seed ORF' is used to describe the minimal, most reliable possible ORF that can be obtained by extending the reliably aligned regions in the upstream direction until the first start codon occurs and in the downstream direction until a stop codon is encountered.

*** Coding potential and the complete set of candidate ORFs.** Seed ORF sets then were utilized to calculate the codon usage tables and the average and standard deviation of the coding potential. Let *F(abc)* be the genomic frequency of the codon *abc*, thus the *primary coding potential* of a DNA segment of length n codons is:

$$Q(a_1b_1c_1 \ldots a_nb_nc_n) = \sum_{k=1}^{n} \log F(a_kb_kc_k)$$

Then the normalized potential measured in the standard deviation units were used to account for DNA fragments of different length, and the coding quality of a DNA fragment is defined. Upon derivation of the statistical parameters above, the DPS output was screened again to extract all similarity-based seed ORFs. All seed ORFs were then extended in the 5' direction as far as possible. Short overlaps between genes were allowed. Each extension piece of DNA started with ATG, GTG or TTG. This procedure resulted in the complete set of 'open-start' candidate ORFs.

*** 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.

Actually, a part of RBS is formed by the purine-rich Shine-Dalgano (SD) sequence which is complementary to the 3’ end of the 16S rRNA. Let *F(b, j)* be positional nucleotide frequencies in the initial alignment *[j=(-20) ... (-1); b= T, C, A, G]*. Then positional information content is defined as:

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

G(b) is the genomic frequency of the base b. Initially the RBS signal was assumed to reside in positions having the maximum total information content, then the position of the SD box in each individual sequence was determined using the two step iterative procedure, involving selection of top scoring (SD signal score, usually top 80%) fraction of sequences at each optimization step; the preference for the distance between the SD box and the start codon are taken into account. Thus the RBS profile is the nucleotide weight matrix and the vector of position weights. Finally the genome annotation is the assignment of start of 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. That is, in ORFs with multiple candidate starts, the leftmost start codon having sufficiently strong RBS is selected.

III Results and discussion

529 *B. subtilis* and 910 *E.coli* similarity-based seed ORFs were extracted from the DPS search results. Figure 1 is the alignment of the *B.subtilis* regions upstream from the 5’ ends of the ORFs with one possible start codon and acceptable coding potential. In table I, RBS weight matrices were derived from alignments of 385 *B. subtilis* and 644 *E.coli* 5’ upstream gene regions with single candidate starts (as in Figure 1).

In Figure 2, certain locations of candidate SD box relative to the start codon, in this case –13 both in *B. subtilis* and 910 *E.coli*, are strongly preferred. Thus incorporation of the positional preference information in addition to the standard weight matrices allows for higher selectivity in RBS detection.

Figure 3 illustrates the search for an appropriate start position of the *B. subtilis comA* gene after its seed ORF staring in position 3252280 of the genome was detected. As seen, the ATG start codon in position 3252523 was accepted since it has a very strong RBS upstream. This corresponds to the gene start position identified by Weinrauch *et. al.* [5].

4379 genes longer than 35 codons in the *B. subtilis* genome and 4595 genes longer than 35 codons in the *E.coli* genome were identified in this work. In Figure 4 we can see that most of the false negatives (genome proteins not identified) and false positives (overpredicted ORFs) are shorter or slightly longer than 100 condons. The number 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. But for shorter ORFs, the prediction deteriorates. Note that the main source of errors in the length 20-80 codons is false positives, or say, unsupported ORFs for which no experimental evidence proving their existence is available, and can not be excluded because some, or even many of these putative genes may be real.

Table 2 gives the comparison of the gene prediction results with the sets of sequences from the PIR-international (SUBPIR and ECOPIR) and the genome sequencing projects (SUBGEN and ECOGEN). Gene and gene star predictions were very accurate in this work. 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 can be explained by the details of the gene analysis in the original publications. Also, 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 [6]; 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 [7].

The main distinctive feature of this algorithm is that to 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. Unlike GeneMark and EcoParse, the algorithm does not rely on the statistics of the non-coding regions (only coding regions can be defined unambiguously, especially at the initial steps of the analysis).Not using the energy of the base-pairing of the SD and the 16s rRNA makes the program applicable, either to early stages of genome analysis when the rRNA may be not be sequenced yet, or to the situation when RBS does not conform to the standard base-pairing model. Finally, instead of using complicated multiple alignment techniques for derivation of the RBS profile, this algorithm use a relatively simple iterative procedure to detect the relatively strong RBS signal.

For better prediction of genes and their starting positions, additional considerations may be taken into account, such as the influence of the mRNA secondary structure on the choice of start codons. Protein features can also be important for gene recognition.

[1] D. Frishman, *et. al.,* Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, *Nucleic Acids Research*, 26(12), 2941-2947, 1998
[2] T. Etzold, *et. al., Methods of Enzymol.,* 266, 114-128, 1996
[3] W. Gish, *et. al., Nature Genet*, 3, 266-272, 1993
[4] X. Huang, *Microb. Compar. Genomics,* 1, 281-291, 1996
[5] Y. Weinrauch, *et. al., J. Bacteriol.,* 171, 5362-5375, 1989
[6] D. C. Shields, *et. al., Nucleic Acids Res.,* 15, 8023-8040, 1987
[7] A. G. Hatzigeorgiou, et. al., Locating translation initiating sites in Bacillus subtilis and other Gram-positive bacteria, First Annual Conference on Computational Genomics, p.8., 1997

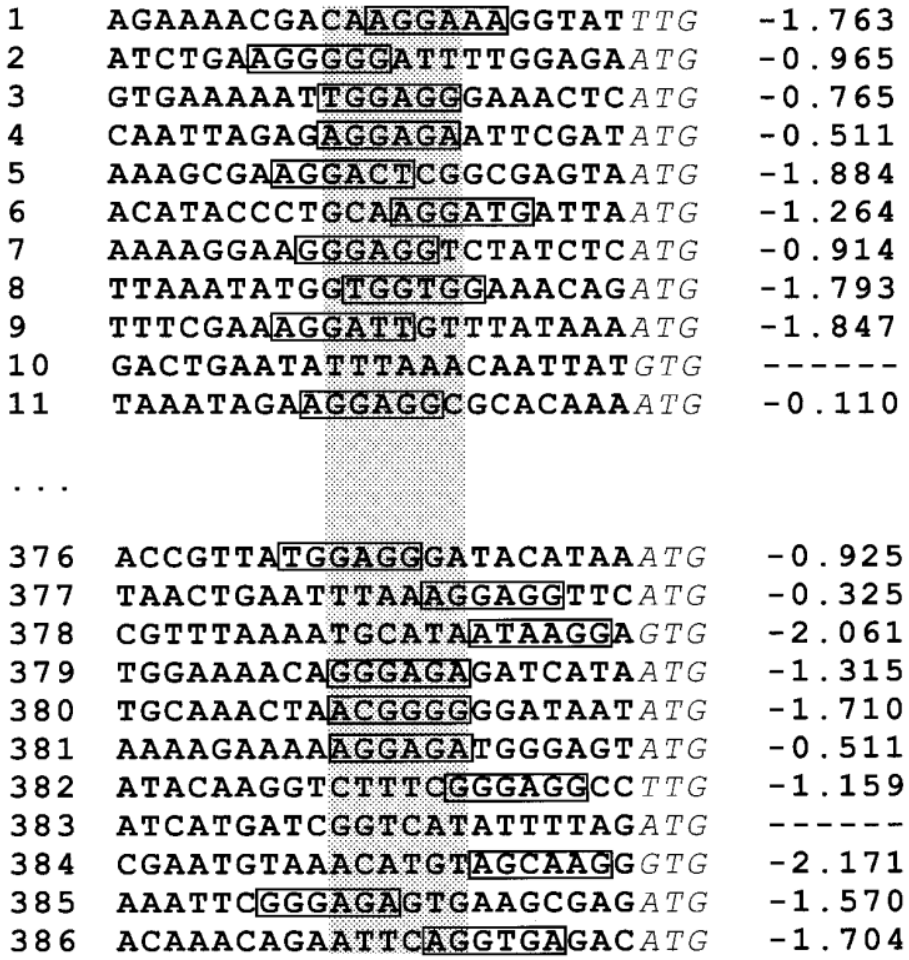
Table 1. RBS weight matrix for *B.subtilis* and *E.coli*

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

Table 2. caption>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.



1	AGAAAACGACAAGGAAAGGTAT*TTG*	-1.763
2	ATCTGAAGGGGGATTTTGGAGA*ATG*	-0.965
3	GTGAAAAATTGGAGGGAAACTC*ATG*	-0.765
4	CAATTAGAGAGGAGAATTCGAT*ATG*	-0.511
5	AAAGCGAAGGACTCGGCGAGTA*ATG*	-1.884
6	ACATACCCTGCAAGGATGATTA*ATG*	-1.264
7	AAAAGGAAGGGAGGTCTATCTC*ATG*	-0.914
8	TTAAATATGGTGGTGGAAACAG*ATG*	-1.793
9	TTTCGAAAGGATTGTTTATAAA*ATG*	-1.847
10	GACTGAATATTTAAACAATTAT*GTG*	------
11	TAAATAGAAGGAGGCGCACAAA*ATG*	-0.110
...		
376	ACCGTTATGGAGGGATACATAA*ATG*	-0.925
377	TAACTGAATTTAAAGGAGGTTC*ATG*	-0.325
378	CGTTTAAAATGCATAATAAGGA*GTG*	-2.061
379	TGGAAAACAGGGAGAGATCATA*ATG*	-1.315
380	TGCAAACTAACGGGGGGATAAT*ATG*	-1.710
381	AAAAGAAAAAGGAGATGGGAGT*ATG*	-0.511
382	ATACAAGGTCTTTCGGGAGGCC*TTG*	-1.159
383	ATCATGATCGGTCATATTTTAG*ATG*	------
384	CGAATGTAAACATGTAGCAAGG*GTG*	-2.171
385	AAATTCGGGAGAGTGAAGCGAG*ATG*	-1.570
386	ACAAACAGAATTCAGGTGAGAC*ATG*	-1.704

Figure 1. Alignment of the *B.subtilis* regions upstream from the 5[prime] 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.

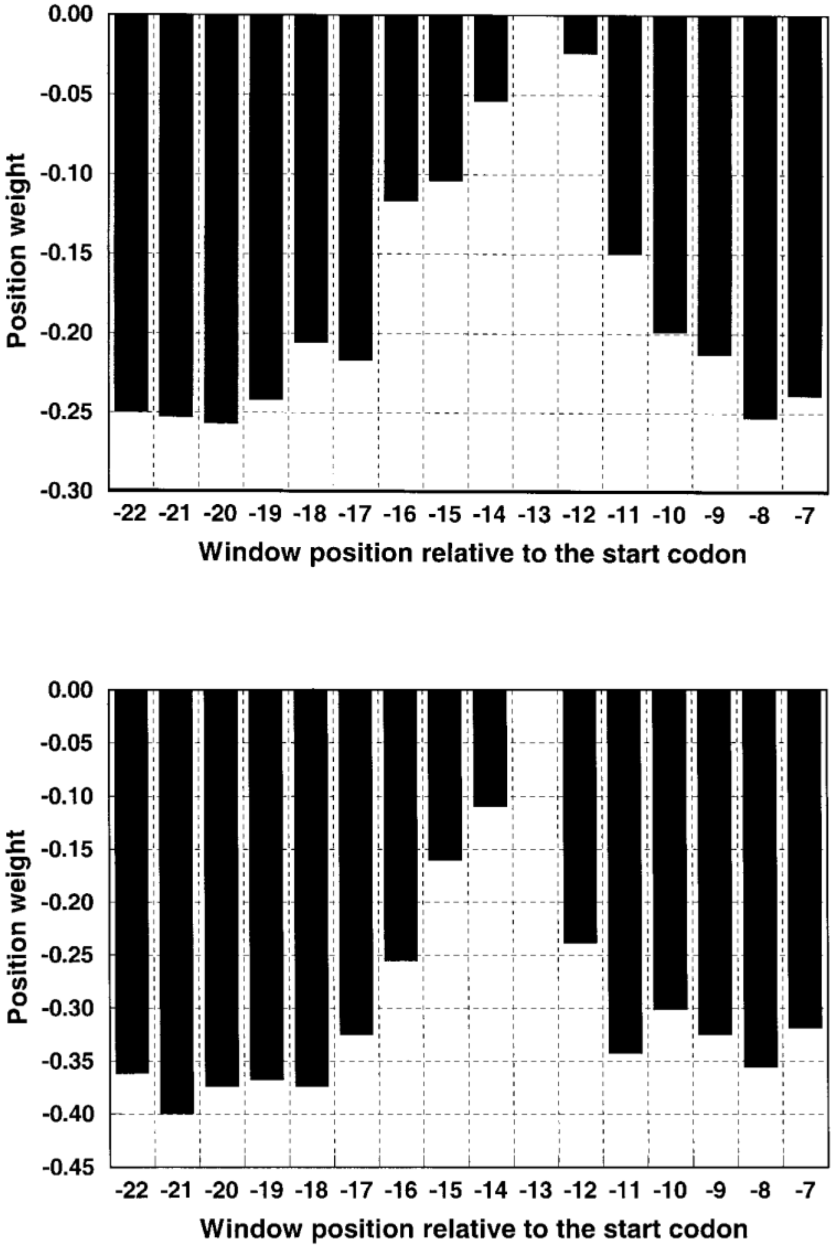
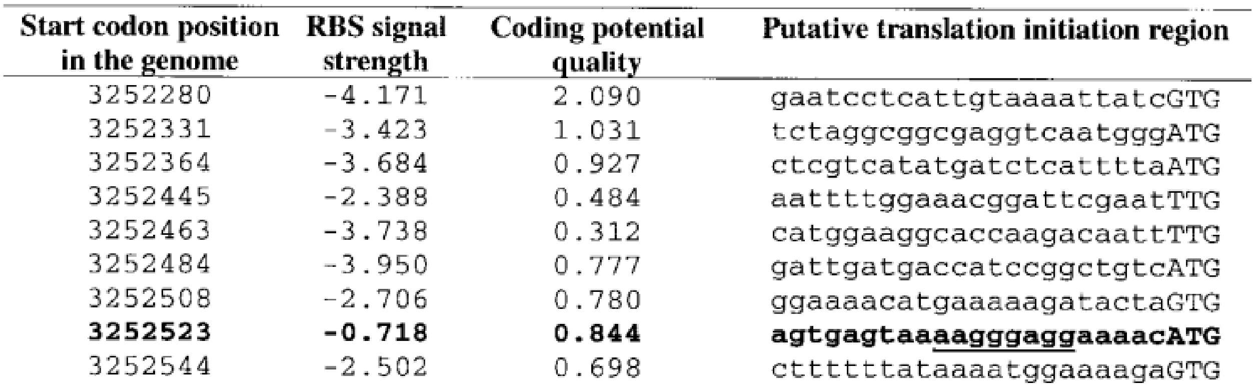


Figure 2. Automatically derived positional preference of the SD box in *B.subtilis* (**a**) and *E.coli* (**b**). Higher values (closer to 0) correspond to more preferred start positions of the window of length 6 nt.



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 ([Omega]) and RBS signal strength upstream ([Delta]) 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 [Omega] are higher than the conservative threshold -1.0 in all cases.

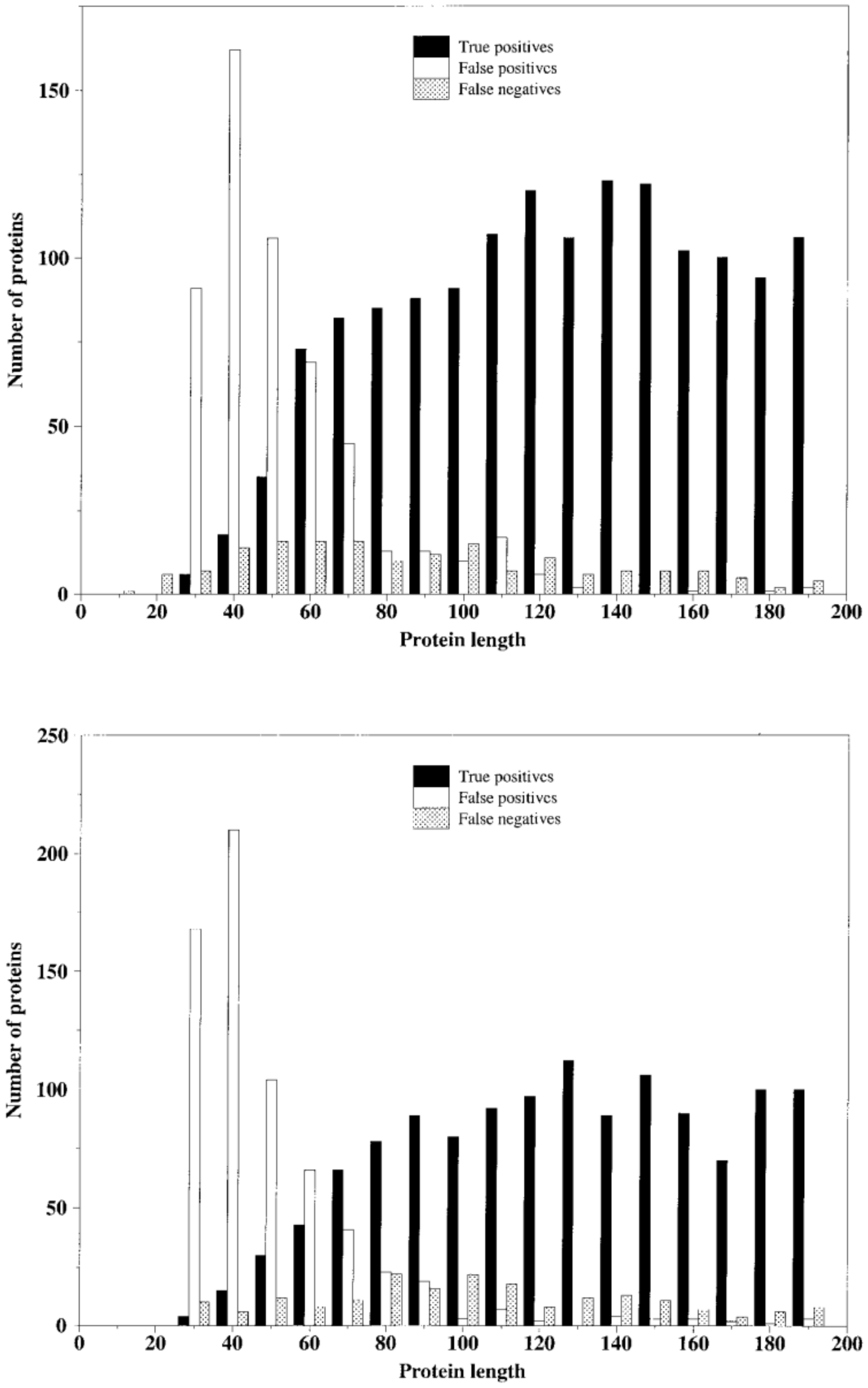


Figure 4**.** Distribution of the true positive, false positive and false negative lengths in *B.subtilis* (**a**) and *E.coli* (**b**) ORFs. Only the ORF length range 0-200 codons is shown. Predictions for longer ORFs are practically perfect.